# Peer Review File

**Manuscript Title:** Quasi-experimental evaluation of a nation-wide lifestyle intervention program for prediabetes in the UK

## Reviewer Comments & Author Rebuttals

### Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

Interesting analysis on the causal effect of lifestyle interventions.

I have few comments for your consideration.

Patients below the threshold and referred to treatment may be different than patients above threshold and not referred to treatment. Describing these groups in detail would be important. I think one of the major confounders not accounted for here is individual/personal motivation. I think the effect of treatment that is measured here likely not only include treatment but willingness of the patient to be referred to such a program. Willingness to be referred can influence willingness and ability to adhere to program guidance and hence improved clinical outcomes. i think the authors should consider this a bit more carefully. The findings almost certainly do not only reflect the causal effect of intervention.

Patients below the threshold who also received treatment may be different. Why would a GP refer them to treatment when their HBA1C is below the threshold? How was this dealt with analytically?

I am also concerned about incomplete accounting for confounding bias. Regression Discontinuity does not completely eliminate confounding. Testing of negative exposure controls (using the same design), and negative outcome controls (using the approach outlined by Lipsitch may be helpful) https://pubmed.ncbi.nlm.nih.gov/20335814/

I really like the causal thinking here, but i am a bit disappointed to see that the authors rely on a sole analytic approach (regression discontinuity) to derive their conclusions. The paper would be strengthened substantially with triangulation of evidence from multiple approaches perhaps formal difference in differences analyses, a trial emulation approach (Miguel Hernan), and other methods of causal inference. Reproducing the findings using multiple approaches (triangulation) would strengthen this considerably. This is important. i really dont think that reliance on one method is enough to validate the conclusions presented here.

Referee #2 (Remarks to the Author):

Strengths

- The manuscript addresses a significant issue and provides a novel, innovative design (regression discontinuity analysis) to evaluate the real-world effectiveness of the Diabetes Prevention Program at a population health level.
- The authors utilize a high-quality data source to evaluate population-level effects.
- The methods are explained in detail and the authors adequately address issues of missing data in large electronic medical record sources and include relevant sensitivity analyses.

Weaknesses

- The authors state they aim to "determine the health effects of the largest behavior change program for pre-diabetes globally" but they evaluate the effect of program referral rather than program participation/utilization. Program referral often does not accurately reflect program enrollment, engagement, and retention, particularly for longitudinal behavior change interventions like the DPP.
- The authors do not account for important confounders that need to be considered if using program referral as a marker of program participation such as duration of prediabetes, previous attempts at lifestyle modification, readiness for behavior change, etc.; these variables would need to be considered since they could influence likelihood of receiving a referral, activation to participate in the program after referral, and/or likelihood of engaging in behavior change independent of program participation.
- More information is needed about DPP availability for the entire population, provider referrals to the program (and predictors of referral) at a population level, and adherence to the 12-month program at a population level, in the period after the guideline change to understand if/how program referral can be used as a marker of program utilization by participants.
- The study design (i.e., use of program referral as marker of participation, duration of follow-up period, etc.) does not account for the 1-year duration of the DPP and the importance of long-term (at least 12 months) sustainability of lifestyle change (dealing with relapse, regain, etc.) which is a key component of the program and predictor of long-term benefits of the program. The effects measured in this study may be more reflective of benefits of identifying individuals with prediabetes rather than the effects of participation in a 12-month program.

**Author Rebuttals to Initial Comments:**

| Number | Comment | Response and Location of Modifications in Revised Manuscript |
|---|---|---|
| Reviewer #1 | | |
| General comment | Interesting analysis on the causal effect of lifestyle interventions. | We are very grateful for the thoughtful suggestions on how to improve our manuscript. We have now carried out a comprehensive revision, which includes **a series of new additional analyses**. In addition to our regression discontinuity approach, we now also show results from difference-in-differences, instrumental variable, propensity-score matching, and multivariable regression analyses, which all confirm the beneficial effects of the NHS Diabetes Prevention Programme (NHS DPP). We have also **added an extensive set of analyses to test for the presence of confounding** bias, as suggested by the reviewer, using the approach outlined by Marc Lipsitch[1]. This series of robustness analyses demonstrates that systematic confounding bias is unlikely in our study.<br><br>We respond to each of the reviewer's comments in detail below.<br><br>1. Lipsitch, M., Tchetgen Tchetgen, E. & Cohen, T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology* **21**, 383–388 (2010). |
| 1 | Patients below the threshold and referred to treatment may be different than patients above threshold and not referred to treatment. Describing these groups in detail would be important. I think one of the major confounders not accounted for here is individual/personal motivation. I think the effect of treatment that is measured here likely not only include treatment but willingness of the patient to be referred to such a program. Willingness to be referred can influence willingness and ability to adhere to program guidance and hence improved clinical outcomes. I think the authors should consider this a bit more carefully. The findings almost certainly do not only reflect the causal effect of intervention. | We thank the reviewer for this important remark. To address these points, we have now added a series of analyses that show that (a) patients below and above the threshold are similar on important characteristics including variables that are a proxy for personal motivation, and (b) patients who are referred and start the program are comparable to patients who are referred and do not start the program. In addition, (c) we have clarified throughout the manuscript that our primary analysis establishes the causal effect of *referral* to intensive lifestyle counseling, and (d) have added a series of analyses to also estimate the effect of *uptake* of the intensive lifestyle counseling program. Lastly, e) we have conducted new analyses (a difference-in-differences as well as an instrumental variable design) that compare entire GP practices against each other (instead of relying on the HbA1c threshold for prediabetes for causal effect estimation) at different time-region combinations of program rollout. In the following, we explain all changes in detail.<br><br>a) <u>Comparing patients below and above the threshold:</u> We agree that it is essential to describe patients below and above the threshold in detail. In a regression discontinuity analysis, the causal effect of being referred due to the threshold rule is not affected by group differences between those below the threshold and referred versus those above the threshold and not referred, as long as patient characteristics that drive the physician's decision to refer a patient to the program are continuously distributed near the threshold[1]. In other words, as long as crossing the threshold causes a sudden jump in only the probability of being referred to intensive lifestyle counseling, but not in any other patient characteristics, regression discontinuity provides an unconfounded effect estimate. Another way to express this rationale is to allude to a clinical trial, in which there is imperfect adherence to the intervention. In this scenario, the intent-to-treat effect is (i.e., the unadjusted comparison of intervention vs control arm) is unbiased despite the imperfect adherence to the intervention (e.g., taking a particular medication). We now demonstrate empirically that crossing the threshold causes a sudden jump in only the probability of being referred to intensive lifestyle counseling through the **addition of a series of balance tests**. These all show that no potential confounders (including demographic characteristics such as age or gender and physiological variables such as baseline body mass index or blood pressure) change abruptly at the threshold. These analyses are shown in Figures 1, S2-5, and further described on (page 5, lines 7-23). |

| | | |
|---|---|---|
| | | In addition, we have now expanded our balance tests to variables that may function as a proxy for health literacy, health services utilization, and personal motivation, such as prior receipt of weight loss counselling, attendance of preventative screening services, and uptake of vaccinations. These are shown in Figures S6-9 and all demonstrate that **none of these variables reflecting personal motivation change abruptly at the threshold**.<br><br>Taken together, our balance tests show that patients below and above the threshold are not systematically different (table S5).<br><br>b) Comparing referred patients who started the program vs. not: We now compare patients who were referred to the NHS DPP and started the intervention with patients who were referred but did not start the intervention. We derived this information from the general practice records and verified their plausibility by comparing the attendance rates with official program reports. Comparison on a multitude of variables shows that **patients who were referred and started the program are similar to patients who were referred and did not start the program** (table S15: all variables yield a standardized mean difference [SMD] below 0.2 and 20 out of 21 variables yield a SMD below 0.1).<br><br>c) Interpretation of effect estimates: We agree that the willingness of a patient to be referred and participate in an intensive lifestyle intervention plays a key role in whether the patient's health improves. We realize that previous language in our manuscript may have been ambiguous, and we have now clarified throughout our manuscript that the primary results presented are explicitly about the causal effect of *being referred* to the NHS DPP and other intensive lifestyle interventions. Specifically, we estimate both the (i) effect of offering the program at its current reach (the effect of the threshold rule) and the (ii) effect of agreeing to being referred (the causal average complier effect). The (i) effect of the threshold rule is valid regardless of the degree of compliance in the sample, as it only measures the effect of the NHS DPP implementation from a policymaker's perspective (i.e., how does offering the program change the outcome, given existing adherence rates?). The (ii) complier average causal effect, in turn, explicitly estimates the causal effect on compliers, i.e., those who accept a referral to the program because they are above the threshold. We believe that it is important for clinicians to know what health benefits they can expect for patients who initially accept an offered referral; at a point where it is impossible for the physician to determine or control whether and to what extent patients will adhere to the program.<br><br>d) Analyses to estimate the effect of intervention uptake: The fact that patients who were referred to the program are similar to those that started the program (as detailed under b) and shown in table S15) allowed us to also approximate the effect of program uptake. This effect approximates the health impact when everyone who is referred actually attends, and may serve as important information for policymakers about potential effects of the program if interventions are undertaken to improve the rate at which patients comply with a referral to intensive lifestyle counseling. We describe this analysis in the paragraph with the sub-header "Intent-to-treat versus per-protocol effect" (page 9, lines 12 ff.).<br><br>e) Difference-in-differences and instrumental variable estimation: We describe these new analyses in detail in response to the last comment by the reviewer. Here, we would simply like to highlight that these new analyses use a different approach for causal effect estimation (leveraging |

| | | |
|---|---|---|
| | | differences in time and space in how the NHS DPP was rolled out) than our regression discontinuity analyses, and thus, measure, a different causal effect and make a different set of assumptions. The fact that our difference-in-differences and instrumental variable analyses (as well as our propensity score matching, multivariate regression, and panel data regression with individual-level fixed effects – also described in more detail in response to the last comment by the reviewer) arrive at the same conclusion as our regression discontinuity approach lends additional confidence that our analysis does not suffer from major confounding.<br><br>1. Lee, D. S. & Lemieux, T. Regression Discontinuity Designs in Economics. *Journal of Economic Literature* **48**, 281–355 (2010). |
| 2 | Patients below the threshold who also received treatment may be different. Why would a GP refer them to treatment when their HBA1C is below the threshold? How was this dealt with analytically? | Patients below the threshold should not have been referred according to public health guidance and program stipulations[1-2]. We have verified with the NHS DPP team from the National Diabetes Audit England that **patients without any HbA1c record ≥ 42 mmol/mol in their general practice records are not accepted into the NHS DPP by any provider** (unless they have a history of gestational diabetes). However, patients may have been referred to an intensive lifestyle intervention that we considered to be comparable to the NHS DPP (e.g., a structured weight loss program) at the discretion of their primary care physician (e.g., due to other risk factors such as excess weight) or based on a second HbA1c test that exceeded the threshold within the referral window of 12 months. Thus, we have now conducted a **robustness analysis in which we only consider referrals to the NHS DPP as the treatment variable** and restrict the referral window to three months after the baseline HbA1c measurement. This led to a sample in which less than three percent (620/20,963) of patients who were referred to the NHS DPP had an HbA1c below the threshold (i.e., < 42 mmol/mol). Encouragingly, the treatment effects in this robustness analysis were very similar to those obtained in our primary analysis (as detailed on page 9, lines 1-10 in our manuscript; and tables S13-14 in the supplement).<br><br>Although we think it is worthwhile to have done this robustness analysis, we would like to point out that strict adherence to the threshold (i.e., all patients above the HbA1c threshold are referred and no patients below the threshold are referred) is not required for our regression discontinuity approach to yield unbiased causal effects. We merely require that there is a sudden change in the probability of being referred at the HbA1c threshold (but this change does not need to go from 0% to 100%). Below, we list several references that discuss this point in more detail[3-5]. Our study design determines the unbiased causal effect of being referred (due to the threshold rule) as long as (a) HbA1c measurements are not manipulated and (b) patient characteristics that may influence the physician's decision to refer a patient to the program do not suddenly change at the threshold. Manipulating a patient's HbA1c measurement is virtually impossible in our scenario given that the exact blood test result is automatically uploaded into the electronic health data system (rather than entered by physicians themselves). This is substantiated by our analyses finding no evidence of heaping in the histogram and density plots of baseline HbA1c values. As explained in more detail in our response above, we also precluded that there is any difference between patients below and above the threshold in a series of balance tests (i.e., we demonstrate visually and empirically that there is no sudden change in relevant characteristics at the threshold).<br><br>1. NICE (National Institute for Health and Care Exellence). Type 2 diabetes: prevention in people at high risk. |

| | | |
|---|---|---|
| | | 2. NHS Health Check - Healthier You: NHS Diabetes Prevention Programme. https://www.healthcheck.nhs.uk/commissioners-and-providers/delivery/healthier-you-nhs-diabetes-prevention-programme/.<br>3. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* **91**, 444–455 (1996).<br>4. Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L. & Bärnighausen, T. Regression Discontinuity Designs in Epidemiology: Causal Inference Without Randomized Trials. *Epidemiology* **25**, 729–737 (2014).<br>5. Oldenburg, C. E., Moscoe, E. & Bärnighausen, T. Regression Discontinuity for Causal Effect Estimation in Epidemiology. *Curr Epidemiol Rep* **3**, 233–241 (2016). |
| 3 | I am also concerned about incomplete accounting for confounding bias. Regression Discontinuity does not completely eliminate confounding. Testing of negative exposure controls (using the same design), and negative outcome controls (using the approach outlined by Lipsitch may be helpful) https://pubmed.ncbi.nlm.nih.gov/20335814/ | Although we would argue that regression discontinuity indeed eliminates confounding when all assumptions are met (and we carefully check each assumption empirically in our manuscript), we entirely agree that explicitly testing negative exposure controls and negative outcome controls is an excellent suggestion that would provide further reassurance of unconfoundedness. We have now implemented a set of analyses using the approach outlined by Lipsitch and colleagues suggested by the reviewer[1], as detailed below.<br><br>a) <u>Negative exposure controls:</u> We created a negative exposure control scenario, i.e., we mimicked an analysis as close to possible to our primary study design without the 'essential ingredient'—the NHS DPP program. To do so, we used an alternative patient cohort that is identical in inclusion criteria to our primary analysis but who had baseline HbA1c measures in 2014 and 2015 – a period during which patients were already identified with prediabetes based on the same threshold and public health guidance. However, the key difference was that during this time period there were very limited intensive lifestyle programs in place to which physicians could refer their patients. This led to a scenario where some key potential confounders (e.g., higher likelihood of being identified with prediabetes as evidenced by a significant increase in the number of primary care consultations) could already be observed, while the hypothesized causal mechanism—the NHS DPP—was 'deactivated', thus constituting a 'placebo analysis' where no effects on outcomes should be observed. Results confirmed that while we indeed observe discontinuities in outcome availability and number of consultations similar to our primary analysis (fig. S11-12), there is no meaningful discontinuity in referrals to intensive lifestyle counseling in 2014 and 2015 (fig. S13) and, most importantly, **no improvement in HbA1c and BMI** (fig. S14-15).<br><br>b) <u>Negative outcome controls:</u> We also tested a set of negative outcome control variables, i.e., health outcomes for which we assumed that any causal effect through intensive lifestyle counselling (our hypothesized mechanism) would be implausible. Those variables are the probability of recording flu vaccination, pneumococcal vaccination, fracture, accidental injuries, routine cancer screenings (bowel, breast, cervical), screening for skin abnormalities, asthma attacks, allergic reactions, hay fever, use of oral contraceptives, use of an intrauterine device (IUD), and short-term onset of cancer or dementia. We repeated the main regression discontinuity analysis for these outcomes. All **results demonstrated no systematic confounding in our primary and secondary cohort** (table S19).<br><br>1. Lipsitch, M., Tchetgen Tchetgen, E. & Cohen, T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology* **21**, 383–388 (2010). |

| | | |
|---|---|---|
| 4 | I really like the causal thinking here, but i am a bit disappointed to see that the authors rely on a sole analytic approach (regression discontinuity) to derive their conclusions. The paper would be strengthened substantially with triangulation of evidence from multiple approaches perhaps formal difference in differences analyses, a trial emulation approach (Miguel Hernan), and other methods of causal inference. Reproducing the findings using multiple approaches (triangulation) would strengthen this considerably. This is important. I really don't think that reliance on one method is enough to validate the conclusions presented here. | We appreciate the reviewer's acknowledgment of the causal thinking that we applied to our analysis. While we believe that regression discontinuity is the strongest causal method that can be applied to our data at hand, we have undertaken extensive efforts to add a series of new analyses that use other approaches for causal effect estimation. We now triangulate the evidence from our regression discontinuity approach by (a) performing a **difference-in-differences analysis** using data before and after the phased roll-out of the NHS DPP program, (b) using the regional variation in program coverage as an **instrumental variable** for program referral, (c) optimizing our cohort based on a **trial emulation** framework, (d) performing **propensity score matching,** (e) conducting **multivariate regression,** and (f) running a **panel data regression with fixed effects. All analyses lead to the same conclusions as our regression discontinuity approach.** We detail all these new analyses below.<br><br>a) <u>Difference-in-differences analysis:</u> We have now performed a difference-in-differences analysis that leverages the phased national roll-out of the NHS DPP program (1st wave start date: 1st June 2016, 2nd wave start state: 1st April 2017, and from 3rd wave start date: 1st April 2018). Specifically, it is possible to compare patients from practices belonging to waves 1 and 2 with patients from practices belonging to wave 3 as the control group. To allow us to aggregate the multiple observed timepoints, we apply the framework by Callaway and Sant'Anna[1]. We describe our methodology in detail in the Methods (page 29, lines 11 ff.). Our estimate of interest, thus, represents how much HbA1c levels in intervention practices (wave 1 and wave 2) have changed in the period after the NHS DPP roll-out, compared to what would have happened in the absence of the program. We describe our results on page 13 in lines 24 ff. Importantly, in the post-DPP introduction period, patients in intervention practices have statistically significantly lower HbA1c levels (Figure 3, table S87). Our difference-in-differences analysis, thus, corroborates our regression discontinuity-based findings.<br><br>b) <u>Instrumental variable estimation:</u> To further triangulate the effect of the NHS DPP on glycemic control (our primary outcome), we present an analysis that leverages regional differences in NHS DPP coverage as an instrument for actual referral to the program. For this, we computed the regional variation in NHS DPP implementation (which varies by program wave) based on official information from the National Diabetes Audit and used a panel dataset of patients with HbA1c records in the years 2017 to 2020 (Methods, page 30, lines 30 ff.). Using panel data, we then computed treatment effects from a Two-Stage Least-Squares regression using the time-variant regional NHS DPP coverage as instrument and controlling for practice- and year-fixed effects. Specifically, our analysis estimates the effect of NHS DPP referral on glycemic control by comparing those who changed from not being referred to being referred due to a change in program coverage in their region. While this estimation approach is relatively imprecise because it relies on time by region variation instead of patient-level variation, we continue to find a significant beneficial effect of program referral on endline HbA1c (page 14, lines 4-10; table S88).<br><br>c) <u>Trial emulation approach:</u> We have now refined our analysis and framing using a trial emulation approach (table S1). The application of the trial emulation approach – a useful lens to make causal thinking comparable across analyses – led to an optimized cohort selection, leaving us with a primary and secondary cohort. Our primary cohort includes patients who had an HbA1c test in a close bandwidth around 42 mmol/mol with no prior use of diabetes medication and no history of prediabetes or diabetes. This closely mimics the cohort for an ideal target trial. In addition, in our secondary cohort, we excluded patients who had any records of previous HbA1c testing to create a |

| | | |
|---|---|---|
| | | more homogenous cohort in which patients are likely to have received very limited lifestyle counseling prior to the intensive lifestyle counseling program. We applied our primary analytic approach (regression discontinuity analysis) to both cohorts and substantially extended our sensitivity analyses (see previous comment). The results were robust across cohorts with the secondary cohort benefitting slightly more compared to the primary cohort. The target trial specifications also informed our newly added regression-based analyses described in the next paragraph.<br><br>d-f) <u>Propensity score matching, multivariate regression, and panel data regression with individual-level fixed effects:</u> We now also include a set of additional regression-based analyses. Based on our target trial protocol, all of these analyses were restricted to patients eligible for the NHS DPP, i.e., adults who had a first HbA1c test ≥ 42 mmol/mol and < 48 mmol/mol with no prior use of diabetes medication and no history of diabetes. First, we conducted propensity score matching (using age, gender, baseline HbA1c, average number of GP consults in previous three years, the index of multiple deprivation [IMD] derived from patients' zip codes, region of GP practice, whether or not patients received prior lifestyle advice, time to endline, and diabetes medication) followed by g-computation to estimate the average marginal effect of NHS DPP referral on HbA1c accounting for confounding by the included covariates[2-3]. We used 1:1 nearest neighbor propensity score matching without replacement with a propensity score estimated using logistic regression of the treatment on the covariates. Second, we computed multivariate regressions adjusting for (i) baseline confounders, i.e., age, gender, IMD, HbA1c, BMI, average number of GP consults in previous three years, receipt of flu or pneumococcal vaccination in previous five years, prior lifestyle advice or weight loss counselling, prior consultation for skin abnormalities, prior participation in bowel cancer screening and (ii) additional post-baseline confounders, i.e., diabetes medication and time to follow-up. Third, using the same panel dataset as for the instrumental variable estimation described above, we calculated the one-year lagged treatment effect of program referral on HbA1c from a regression with individual-fixed effects to account for time-invariant differences between patients and year-fixed effects to control for the secular variation. While these analyses provide, in our view, less robust causal effect estimates than our regression discontinuity approach, all analyses yielded the same conclusions – that is, a significant negative association of being referred to the NHS DPP with endline HbA1c (table S88).<br>References<br>1. Callaway, B. & Sant'Anna, P. H. C. Difference-in-Differences with multiple time periods. *J. Econom.* **225**, 200–230 (2021)<br>2. Snowden, J. M., Rose, S. & Mortimer, K. M. Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique. *American Journal of Epidemiology* **173**, 731–738 (2011).<br>3. Greifer, N. & Stuart, E. A. Choosing the Estimand When Matching or Weighting in Observational Studies. (2021) doi:10.48550/ARXIV.2106.10577. |
| Reviewer #2 | | |
| General | The manuscript addresses a significant issue and provides a novel, innovative design (regression discontinuity analysis) to evaluate the real-world effectiveness of the Diabetes Prevention Program at a population health level. | We thank the reviewer for their review of our manuscript. We are very grateful for the thoughtful suggestions on how to improve our manuscript. In response, we have expanded our analyses to examine the role of, and account for, a patient's motivation in treatment effectiveness. We have also taken the reviewer's concern into account regarding the interpretation of our measured causal effect estimates and have (a) clarified our language throughout the manuscript, (b) extended our analysis to estimate the effect of taking up the program, (c) added numerous new robustness analyses, and (d) complemented our regression |

| | | |
|---|---|---|
| | - The authors utilize a high-quality data source to evaluate population-level effects.<br>- The methods are explained in detail and the authors adequately address issues of missing data in large electronic medical record sources and include relevant sensitivity analyses. | discontinuity approach with a series of new additional approaches for causal effect estimation (difference-in-differences analysis, instrumental variable estimation, propensity score matching, multivariate regression, and panel data regression with individual-level fixed effects). |
| 1 | The authors state they aim to "determine the health effects of the largest behavior change program for pre-diabetes globally" but they evaluate the effect of program referral rather than program participation/utilization. Program referral often does not accurately reflect program enrollment, engagement, and retention, particularly for longitudinal behavior change interventions like the DPP. | We thank the reviewer for making this important remark. We have now clarified throughout the manuscript what our regression discontinuity-based effect estimates reflect and **added a set of analyses that assess the effect taking up the referral to an intensive lifestyle intervention program** (rather than merely assessing the effect of being referred to the program). In addition, we have now complemented our regression discontinuity analyses with **a set of quasi-experimental approaches (difference-in-differences and instrumental variable analyses) that measure a different causal effect** than our regression discontinuity approach.<br><br>We agree that we have previously not made it sufficiently clear that the primary results presented in our manuscript reflect the causal effect of *being referred* to the NHS DPP and other intensive lifestyle interventions rather than program engagement or retention. We would, however, like to point out that the former is precisely what we are most interested in: the health system-level effectiveness of an intensive lifestyle counseling program implemented at scale in routine care. That is, if a lifestyle intervention program is, for example, so time-intensive that few patients engage with it, then our primary effect measure of interest is not that from taking up or fully adhering to program but rather that from referral to the program. We have now clarified our language at various points throughout our manuscript, for example "We aimed to determine the health effects of being referred to the largest behavior change program for pre-diabetes globally […]" (Abstract), "[…] investigating whether being referred to a routine behavior change program in a national health system has the potential to lead to improvements in key cardiovascular risk factors […]" (Introduction), or "[…] we found causal evidence that referral to intensive lifestyle counseling led to improved glycemic control and reductions in BMI, weight, and triglycerides." (Discussion).<br><br>Nonetheless, **we have now expanded the range of causal effects that we assess**. Specifically, our regression discontinuity-based effect estimates now assess each of the three steps in the cascade from (i) offering the program at its current reach (the effect of the threshold rule), (ii) agreeing to being referred (the causal average complier effect), and (iii) taking up the program (the "per-protocol" effect). Each of these effects is described in more detail below.<br><br>(i) The <u>effect of the threshold rule</u> is valid regardless of the degree of compliance or engagement in the sample, as it only measures the effect of the NHS DPP implementation from a policymaker's perspective (i.e., how does offering the program change the outcome, given existing treatment rates?).<br>(ii) The <u>complier average causal effect</u>, in turn, explicitly estimates the causal effect on those who are referred to an intensive lifestyle intervention program because they are above the threshold. In our view, this estimate most accurately reflects the program's effectiveness. We consider it important that practitioners know what health benefits they can expect for patients who initially accept an offered referral; at a point where it is impossible to control whether and to what extent patients will engage in the program.<br>(iii) The <u>"per-protocol treatment effect"</u> or, in other words, the effect of the program if everyone who is referred actually attends, most closely reflects the program's efficacy. It was possible for us to |

| | | |
|---|---|---|
| | | reasonably approximate this effect since patients who were referred and started the program were not systematically different from patients who were referred but did not start the program. We provide more details in the paragraph with the sub-header "Intent-to-treat versus per-protocol effect" (page 9, lines 4-22) in the manuscript and expand further on our per-protocol analysis in our response to the reviewer's third comment.<br><br>In addition, we have conducted **extensive new quasi-experimental approaches that measure a different causal effect** (and under a different set of assumptions) than regression discontinuity.<br><br>a) Difference-in-differences analysis: We conducted a difference-in-differences analysis utilizing the phased national roll-out of the NHS DPP program, which started on 1st June 2016 for the first wave, 1st April 2017 for the second wave, and 1st April 2018 for the third wave. Our approach involved comparing patients from GP practices in the first and second waves with patients from practices in the third wave as the control group. Details of our methodology can be found in the Methods section (page 29, lines 11 ff.). Our primary estimation assesses the change in HbA1c levels among intervention practices (wave 1 and wave 2) after the NHS DPP roll-out, relative to what would have occurred in the absence of the program (wave 3 practices). Our results (page 13, 24 ff.) lead to the same conclusion as our regression discontinuity analyses.<br><br>b) Instrumental variable estimation: We conducted analyses that utilized regional differences in program coverage as an instrument for actual referral to the program. We calculated the variation in NHS DPP implementation across regions, which varies by program wave, using information from the National Diabetes Audit and a longitudinal dataset of patients with HbA1c records between 2017 and 2020 (Methods, page 30, lines 30 ff.). We then used panel data to compute treatment effects through a Two-Stage Least-Squares regression that controlled for practice- and year-fixed effects, and used time-variant regional NHS DPP coverage as an instrument. As for our difference-in-differences approach, these instrumental variable analyses confirmed the findings from our regression discontinuities (see page 14, lines 4-10; table S88). |
| 2 | The authors do not account for important confounders that need to be considered if using program referral as a marker of program participation such as duration of prediabetes, previous attempts at lifestyle modification, readiness for behavior change, etc.; these variables would need to be considered since they could influence likelihood of receiving a referral, activation to participate in the program after referral, and/or likelihood of engaging in behavior change independent of program participation. | We thank the reviewer for this thoughtful comment. First of all, we would like to clarify that our primary interest is not in using program referral as a marker for program participation. Rather, our analysis primarily aims to assess the causal effects of offering an intensive lifestyle intervention program to patients, as we believe that this is of highest interest to both health system-level policymakers and clinicians (who may not know beforehand the degree to which a particular patient will adhere to the program). We have now clarified this point throughout the manuscript.<br><br>Nonetheless, we agree with the reviewer that useful additional information can be obtained from assessing the causal effect of actually participating in a lifestyle intervention program. As detailed in our response to the preceding comment, we have now added such a "per-protocol" analysis to our manuscript. For this analysis, it is important to know the role of the potential confounding variables listed by the reviewer. We now examine this issue in detail by (a) **refining our cohort selection**, (b) **extending our balance tests** to include a series of variables that are a proxy for health literacy and personal motivation, (c) **adjusting our main regression for these variables,** (d) **stratifying our analyses** by these variables, and (e) **comparing patient characteristics of those who started the program vs. not.** We detail these new analyses below. |

a) Refining our cohort: Following a trial emulation approach, we have now specified our population of interest such that we only include patients who do not have a history of prediabetes. We do so because it is likely that patients who have been previously diagnosed with prediabetes (and continue to have prediabetes) respond differently to being referred to intensive lifestyle counselling. In secondary analyses, we also examine the effect of intensive lifestyle counseling among a cohort of patients that includes both those with new and previously diagnosed prediabetes.

b) Extending our balance tests: We have extended our balance tests to include variables that are a proxy for health literacy and personal motivation. Specifically, we present balance tests for the following variables: previous receipt of lifestyle or weight loss advice, attendance of screening for various forms of cancer (bowel, skin, breast), receiving influenza or pneumococcal vaccination, and adhering to prior statin prescriptions. Balance tests assess whether patients are comparable just below and above the threshold by verifying that variables are continuously distributed near the threshold. As it is in randomized controlled clinical trials, evidence of balance on baseline observables provides confidence that patients assigned to treatment and control conditions are exchangeable, which is a necessary condition for a regression discontinuity analysis to provide valid causal effect estimates. We demonstrate that this condition is met by plotting the relationship between baseline HbA1c and each potential confounder (Figure 1, fig. S2-S9) and computing placebo local linear regressions (table S5). **All balance tests show that patients below and above the threshold are not systematically different.**

c) Adjusting for potential confounding variables: We have expanded our sensitivity analyses by adjusting our regression discontinuity analyses for the following variables: having previously received general lifestyle advice vs. not; having previously received counselling on weight loss vs. not; having previously participated in a skin cancer screening activity vs. not; having previously participated in a bowel cancer screening activity vs. not; and having previously received a flu vaccination vs. not. None of these adjustments altered our results in a significant way (tables S17-18).

d) Stratified analyses: Many of the proxy variables for health literacy and personal motivation listed above (and that we have shown to be equally distributed below and above the threshold in balance tests) may be viewed as *modifier* of the treatment effect rather than as a confounder. Thus, we have now also stratified our analyses by, and provide heterogeneous treatment effect estimations for, the following comparisons: having previously received general lifestyle advice vs. not; having previously received counselling on weight loss vs. not; having previously participated in a skin cancer screening activity vs. not; having previously participated in a bowel cancer screening activity vs. not; having previously received a flu vaccination vs. not; and having previously adhered to their statin prescription vs. not. While these results are exploratory given the absence of pre-specified hypotheses and lower statistical power, the effect estimates suggest that patients who had no record of previously receiving lifestyle advice from their GP benefitted more from being referred to intensive lifestyle counseling. This is also in line with results from our secondary cohort that suggest that patients whose HbA1c were recorded for the first time benefitted more compared to patients whose HbA1c has been previously tested. Taken together, patients' lifestyle modification attempts, health literacy, and personal motivation are likely to play a role in the extent to which program referral ultimately leads to clinical health improvements. All these

| | | |
|---|---|---|
| | | analyses are displayed in tables S57-86 and discussed in section "*Men may be benefitting more than women*", page 12, lines 11-23, and section "*Comparing our "real world" results to effects in clinical trials*", page 15, lines 18-27.<br><br>e) Comparison of patients who started the program vs. not: Lastly, we have compared patients who were referred to the program and started the intervention with patients who were referred to the program but *did not* start the intervention. Specifically, we did not observe any meaningful differences in age, gender, baseline BMI, baseline weight, baseline blood pressure (systolic or diastolic), baseline serum lipids (HDL, LDL, total-to-HDL cholesterol ratio, or triglycerides), number of previous GP consultations, probability of having previously received general lifestyle advice, probability of having previously received counselling on weight loss, probability of having previously participated in a cancer screening activity (skin, bowel, or breast), probability of having previously received routine vaccination (flu or pneumococcal), probability of having previously adhered to their statin prescription, index of multiple deprivation, and rural or urban practice location. We were, thus, able to newly add an analysis that estimates the "per-protocol treatment effect" mimicking a situation where referral can be considered equivalent to program adoption. We detail this new approach more in our response to the next comment.<br><br>Lastly, we would like to highlight that **our new set of quasi-experimental analyses (difference-in-differences and instrumental variable analyses) rely on a different set of assumptions** than our regression discontinuity-based analyses. The finding that these approaches arrive at the same conclusion as our regression discontinuities provides confidence that our analyses do not suffer from major confounding. |
| 3 | More information is needed about DPP availability for the entire population, provider referrals to the program (and predictors of referral) at a population level, and adherence to the 12-month program at a population level, in the period after the guideline change to understand if/how program referral can be used as a marker of program utilization by participants. | As explained in our response to the preceding comments (and as now clarified throughout the manuscript), our primary interest was in determining the causal effect of being referred to an intensive lifestyle counseling program, regardless of the degree of program utilization. Nonetheless, we agree that useful additional information can be obtained from also assessing the causal effect of actually attending the program. **We now (a) provide a synthesis of articles and published reports describing the reach of, and adherence to, the NHS DPP and (b) present what proportion of patients who received a referral actually started the program according to their primary care records.** The latter allowed us to approximate the "per-protocol" effect or, in other words, the effect of the program if everyone who is referred actually attends. We present more detail on these two additions to our manuscript below.<br><br>(a) Reach of, and adherence to, the NHS DPP: We have now added the following information on availability and adherence to the program at a population level in the Methods section (page 24f., lines 21 ff.): "Initial projections by Public Health England and NHS England expected that GP identification would generate demand for around 100,000 participants per year once the program is rolled out nationally by 2020. Official reported statistics from the National Diabetes Audit confirm that referral numbers (offers that were made and not declined) increased steadily from 2017 to 2020, with reporting periods covering January to March of the next year exceeding modelled targets as early as of 2018 (2017: 103 295; 2018: 241 255, 2019: 386 025; and 2020: 559 770)[76]. In a recent analysis from the DIPLOMA research program (Diabetes Prevention – Long Term Multimethod Assessment), it is reported that by April 2020, providers had received a total of 513,312 referrals, of which 271,208 (52.8% of total referrals) had attended an initial assessment and 101,175 (19.7% of total referrals) had attended at least 60% of the program's sessions[75]. Publications report on uptake rate and progression through the program in different ways. |

| | | |
|---|---|---|
| | | According to a scoping review published in early 2022, uptake of the initial assessment following a referral varied from 40% to 78%[77]. Others report on attendance at the first intervention session with rates varying from 36% to 94%[77]. Similarly, numbers on how many participants reached the mid- or endpoint of the program vary considerably. According to an early service evaluation, as of the end of December 2018, 36% attended at least one intervention session and 19% attended at least 60% percent of the sessions[33]. In this early service evaluation, younger age, Asian and Mixed ethnicity (compared to White ethnicity), a lower socioeconomic status, and baseline obesity were associated with lower rates of completion."<br><br>(b) Effect of program uptake: We agree that understanding program efficacy (i.e., the "per-protocol" effect or, in other words, the effect of the program if everyone who is referred actually attends) also provides useful information. It is important to consider that program attendance may be a function of individual characteristics such as personal motivation or available time resources. However, we show that patients who started the treatment did not differ systematically compared to patients who were referred but did not start the treatment, including on the proxy variables that indicate heightened personal motivation or health literacy such as attending preventative cancer screening services or vaccine uptake (table S14-15). We were, thus, able to approximate the effect of program uptake. We did so by first determining whether the patient started the intervention sessions of the NHS DPP program according to GP records. In line with statistics from an official program evaluation[1], among eligible patients that were referred to the NHS DPP program in our primary cohort, 28.1% of patients started the intervention. By scaling up the effect of program referral in the selected bandwidth, we approximate that patients' HbA1c level would be reduced by 3.0 mmol/mol if patients perfectly adhered to referrals and took up the program. While this effect likely represents the upper boundary of the potential population health impact when patients are perfectly compliant with referrals, it may serve as important information for policymakers about the program's maximum potential, for example when deciding about its expansion. We refrained from scaling the effect to fully completing the program since documentation of program completion in GP records is not only likely to be fragmentary but, maybe more importantly, whether or not people are able to fully complete the NHS DPP is in our view part and parcel of its real-world effectiveness. We detail our approach to estimating the effect of program uptake on page 9 f. in lines 12 ff. in the manuscript.<br><br>1. Valabhji, J. *et al.* Early Outcomes From the English National Health Service Diabetes Prevention Programme. *Diabetes Care* **43**, 152–160 (2020). |
| 4 | The study design (i.e., use of program referral as marker of participation, duration of follow-up period, etc.) does not account for the 1-year duration of the DPP and the importance of long-term (at least 12 months) sustainability of lifestyle change (dealing with relapse, regain, etc.) which is a key component of the program and predictor of long-term benefits of the program. The effects measured in this study may be more reflective of benefits of identifying individuals with prediabetes rather | We thank the reviewer for this important remark. As explained in detail in our response to the preceding comments, our primary interest was not in using referral to intensive lifestyle counseling as a marker for participation, but rather to determine the causal effect of the referral itself. We agree that an important concern is that our causal effect estimate may reflect the benefits of identifying individuals with prediabetes rather than the effect of referral to an intensive lifestyle counseling program. We have now conducted **a series of new analyses, which all demonstrate that the presented effects are attributable to the intensive lifestyle counseling program itself.** These new analyses consist of: (a) **"placebo analyses"** showing that there are no health benefits of being identified with prediabetes prior to the roll-out of the NHS DPP program, (b) **two additional quasi-experimental identification strategies**, namely a difference-in-differences analysis and an instrumental variable analysis, that (unlike our regression |

| | | |
|---|---|---|
| | than the effects of participation in a 12-month program. | discontinuity approach) both do not rely on the HbA1c prediabetes threshold for causal effect estimation, and (c) **analyses restricted to patients with a long follow-up period**.<br><br>(a) Placebo analyses: In the years 2014 and 2015, individuals were already identified with prediabetes based on the same threshold and public health guidance as in our study period. Indeed, our data shows that during this period patients were more closely monitored by their physician upon exceeding the threshold (i.e., being identified with prediabetes) as evidenced by a significant increase in the number of GP consultations and shorter follow-up times. However, the key difference compared to our analysis is that there were very limited intensive lifestyle counseling programs in place to which physicians could refer their patients. If our observed health benefits would be attributable to identifying patients with prediabetes (and, for example, the closer monitoring that goes along with such a diagnosis), we would expect to observe similar—or at least some magnitude of—beneficial effects on HbA1c or BMI during the period in which the NHS DPP had not yet existed. However, we did not observe any such health benefits (fig. S14-15). Only after the NHS DPP program was rolled out did we observe significant health benefits in our regression discontinuity analyses. These placebo analyses are presented on page 10 f., lines 23 ff.<br>(b) Difference-in-differences analysis and instrumental variable estimation: These newly added analyses leverage the temporal and regional variation in NHS DPP roll-out, respectively. In both study designs, we compare individuals who were registered in GP practices that had not yet implemented the NHS DPP with patients from GP practices that had already rolled out the NHS DPP. We estimate lagged treatment effects on patients' HbA1c taking the 1-year duration of the NHS DPP into account. We present the results of these analyses on page 12 f., lines 24 ff. The results confirm a significant beneficial effect of the NHS DPP roll-out on glycemic control. Critically, unlike our regression discontinuity approach, these analyses do not rely on the HbA1c threshold that defines prediabetes for causal effect estimation. Thus, the fact that we observe the same health benefits from the NHS DPP program in these analyses is strong evidence that the effects observed in our regression discontinuity approach are due to the NHS DPP program itself (rather than merely due to identifying individuals with prediabetes).<br>(c) Robustness to longer follow-up: The median time between baseline and endline HbA1c measure in our main analysis is 20.5 months. We now present sensitivity analyses to our regression discontinuity analyses in which we impose a lag of at least 12 months or 18 months between baseline and endline measure to account for the duration of the program (as opposed to at least 6 months in the main analysis). These robustness analyses also find significant health benefits from being referred to an intensive lifestyle counseling program (tables S20-S23).<br><br>Nonetheless, although we can confidently exclude the possibility that our causal effect estimates are *entirely* due to identifying individuals with prediabetes, we cannot with absolute certainty exclude the possibility that identifying individuals with prediabetes *partially* contributes to our effect estimates. We, therefore, now clearly state that this is a possible limitation in the discussion section of our manuscript: "Another potential limitation is violation of the exclusion restriction, whereby treatments or exposures other than intensive lifestyle counseling are affected by crossing the cutoff. While we could not precisely control for the relationship between drug dosage and secondary outcomes, we are confident that the observed health effects were not attributable to increased medication following treatment assignment as effect estimates were robust to adjusting for drug prescriptions. Further, it is likely that patients classified as high risk for progression to T2DM have been under closer observation by GPs and that GPs may have taken additional steps to mitigate the risk. We empirically demonstrate that this is unlikely to be a substantial confounder in |

| | | |
|---|---|---|
| | | our analysis because we found no health improvements from crossing the prediabetes threshold prior to NHS DPP roll-out. Similarly, outcomes such as HbA1c or BMI were measured more frequently by GPs in the follow-up period if patients had crossed the threshold. However, testing a diverse set of so-called negative outcome controls—variables that we did not expect to be plausibly affected by referral to the NHS DPP—suggests no systematic confounding. We further conditioned our results on variables that may function as a proxy for health literacy, health services utilization, and personal motivation, which did not substantially change our results, and performed a sensitivity analysis restricting our sample to patients with regular GP visits before their treatment assignment." (Pages 17 f., lines 17 ff.) |

**Reviewer Reports on the First Revision:**

Referees' comments:

Referee #1 (Remarks to the Author):

The authors responded to my comments. they added multiple analyses including a triangulation approach testing different methods which all yielded consistent results. They also responded to my comments to test negative exposure and outcome controls by testing these controls -- and tests were successful.

I have a couple more comments. The finding is interesting, but it really remains very unclear how just referring people to the program yields better outcomes.

a. What remains unclear is how? how just being referred (the act of referral alone) leads to improved outcomes -- certainly this is coupled with post-referral participation in the program, adherence, implementation of advice received, etc.... to then obtain the health outcomes. What is not clear is what is the causal pathway from referral to outcomes? this is important to understand. A DAG would be helpful and analytic approaches to help us deeply understand the causal pathway (and potential mediators) between referral and health outcomes. I worry policy makers will take this, as lets "set it and forget it" lets refer them and not worry about what happens next.

b. in addition to potential causal pathways and mediators, are there any time varying confounding that also need to be fleshed out? or represent threats to validity.

Referee #2 (Remarks to the Author):

-The paper addresses a significant topic and utilize innovative methods.
-The authors made several changes in response to the previous reviewers which strengthened the manuscript. Notably, they have clarified throughout the manuscript that the study evaluates effects of program referral rather than the program itself; they have included stratified analyses to examine potential effect modifiers; they have provided detailed information about DPP implementation; and they have examined effects at a variety of time points after referral.
- A large amount of important information is contained in the supplemental files that cannot be properly discussed in the paper (and cannot fit in a single manuscript). For example, the authors make many general statements about results presented in the supplemental files (for the sake of brevity) that overlook important findings. For example, the authors state patients who referred and started the program are similar to patient who were referred and did not start the program yet table S15 shows some statistically significant differences between the groups. It is also difficult to keep track of primary and secondary cohorts, analyses, etc. in the supplementary files because of the large number of tables and figures; there is not enough space to call out and/or discuss each item in the primary manuscript.

- The authors included a historical control (placebo analyses) to look at the influence of prediabetes diagnosis on clinical outcomes when DPP referral was not available; these analyses are very helpful in showing that prediabetes awareness alone does not lead to the observed outcomes. However, the authors also briefly mention rates of testing for prediabetes increased over time as providers were increasingly able to refer to the DPP (i.e., prediabetes testing increased with increased availability of the DPP for patients found to have prediabetes). This finding deserves more discussion—could increased availability/access to DPP have effects on provider behaviors (screening and referral) and provider/patient perceptions of labeling/diagnosing a precondition during the evaluation period? Could these effects influence the study results?

Referee #3 (Remarks to the Author):

This is a very thoughtful study, with elegant application of several strong quasi-experimental designs, providing internally valid suggestion that referral to diabetes prevention programmes results in improved HbA1c and risk factors but not complications.
The paper is well written, clearly presented, and brings some methodologic diversity to this question. Prior reviews have focused on addressing potential confounding and selection biases, for which these designs can never fully address, but nevertheless I consider the responses of the authors on these issues to be adequate. The additional questions and limitations in the paper are the following:
1. The background introduction is well written in general but ignores the large and growing literature of evaluations of the UK NDPP as well as other settings.
2. The finding of the threshold is clear enough, but it is not clear what assumptions are necessary to conduct the "scaling" of effect size from the eligible-group effect to the referred group effect size. It is not clear why the referred group effect size cannot be calculated based on observed data independent of the eligible denominator. Doesn't this also have to make the assumption that the referred person has no selection effects differentiating them from the eligible but not referred? If there is unmeasured differences there, what would the effect on main findings be?
3. As described, the main finding of benefit is derived from an observation that people in the 6.0-6.4 A1c range saw greater improvement than < 6%. Could the characteristics, meds, health status, changes in risk etc. of the group in the 6.0 to 6.4% category also be affected by changes in the category of them (i.e, diagnosis of diabetes at the 6.5% level)? In other words, if there were shifting diabetes diagnosis, particularly an increased in detection at lower levels than this would pull people from the upper end of the 6.0 to 6.4 range into the diagnosed category, could this result in a spurious decrease in the A1c of the pre-diabetes group?
4. If there is adequate data to assess change in A1c, risk factors, AND complications, its not clear why the paper would not generate an effect size for diabetes incidence. At the same time, the analysis related to complications does not seem particularly useful. If we actually saw a significant effect there, within 1 year, we would be more apt to assume that is reverse causality anyway, with people without complications more likely to be referred, rather than vice versa.

Referee #4 (Remarks to the Author):

As per request from editor I have reviewed the paper from a statistical/causal inference perspective.

* Main remark: I find the statistical analyses included very comprehensive and state-of-the-art. I would like to congratulate the authors on an impressive work.

* Correctness: As far as I can tell the method are correctly applied. They do indeed build an different causal assumptions, which is a strength. However, it does also imply that the methods are not targeting exactly the same parameter. This imply that - in the extreme - we would not require they arrive at the same estimate. Weaker properties such as "same sign and significance" is perhaps all we can demand.

* Use of the multiple methods: The many different methods employed has the natural consequence that the amount of model output is LARGE. This is unavoidable. However, I think the authors could still make one more table or figure (yes I see the irony) which illustrated the estimates on the key outcome obtained from the different methods. Right now it took me ages to find out exactly which estimates should be compared with which. This would also make it easier for the reader to assess if the agree, that the conclusions are indeed supportive.

**Author Rebuttals to First Revision:**

*Referees' comments:*

***Referee #1**:*
*The authors responded to my comments. they added multiple analyses including a triangulation approach testing different methods which all yielded consistent results. They also responded to my comments to test negative exposure and outcome controls by testing these controls -- and tests were successful.*

*I have a couple more comments. The finding is interesting, but it really remains very unclear how just referring people to the program yields better outcomes.*

1. *What remains unclear is how? how just being referred (the act of referral alone) leads to improved outcomes -- certainly this is coupled with post-referral participation in the program, adherence, implementation of advice received, etc.... to then obtain the health outcomes. What is not clear is what is the causal pathway from referral to outcomes? this is important to understand. A DAG would be helpful and analytic approaches to help us deeply understand the causal pathway (and potential mediators) between referral and health outcomes. I worry policy makers will take this, as lets "set it and forget it" lets refer them and not worry about what happens next.*

We agree with the reviewer that understanding the mediating mechanisms such as adherence and behavioral changes adapted by participants is an important research matter. Unfortunately, while routinely available health records have many advantages, one of the drawbacks is that such level of detail is either not reliably documented by general practitioners or not available to us. Thus, while DAGs can be extremely helpful in understanding a causal pathway, we do not see merit in including one in the already very dense manuscript given that we cannot investigate these pathways with the data at hand. **We set out to study the effect of program referrals in an "observational, non-interventional trial in a naturalistic setting" akin to phase IV in drug development** (for which, in fact, adherence is often neglected[1]). As the reviewer has pointed out, our study has consistently demonstrated in a series of quasi-experimental analyses that the presented effects are attributable to the program. We believe that other study types such as randomized control trials or other experimental approaches to mediation such as implicit-mediation analysis[2] are more suited to illuminate mediating factors, for example behavior change techniques that may be chiefly responsible for improvements in physical activity and dietary intake[3] (which in turn is likely to lead to the observed improvements in health outcomes) or adherence to the program. Indeed, early evidence suggests that the number of attended sessions in the NHS DPP acts as a mediator for associations of age, socio-economic status, and ethnicity with weight loss at follow-up[4]. Similarly, higher session attendance mediated the observed effect that participants starting in January lost more weight than those starting in other months[5]. We now briefly discuss the following in the text:

- "Finally, due to the usage of electronic health records we had no detailed information about adherence or persistence to behavior change programs and lifestyle counseling over time. Early evidence suggest that the number of attended sessions in the NHS DPP acts as a mediator for successful weight loss at follow-up[4]. However, our primary interest was in whether the benefits of the NHS DPP are seen in real world circumstances, outside the controlled conditions of a randomized controlled trial. Non-adherence to the program, for example because patients feel that the visits are too frequent or lengthy, are part and parcel of such a real-world effectiveness assessment.", p. 13, ll. 3-8

We share the reviewer's concern that policymakers may limit their efforts to setting up a program such as the NHS DPP and do not consider the important role of adherence (or worse, social inequalities that are exacerbated by differential uptake or adherence). We now include this in the manuscript as follows:

- "[…] While the NHS DPP is operating at a large scale, with 100 000 referrals being offered in 2021[6], there remains a substantial share of adults in England with impaired glycemic control who are presently not taking part in intensive lifestyle counseling, whether it is because of system-level (e.g. placements in the NHS DPP are not available) or physician- and patient-level reasons (e.g. GPs are not compliant with the guidelines). Especially given low uptake and high attrition rates amongst socio-economically disadvantaged and diverse ethnic groups[7], our study lends support to calls for further investment in behavioral interventions and targeted prevention strategies for individuals at risk for diabetes that are currently not reached through care pathways. It should be further investigated whether some groups

may profit from combining programs such as the NHS DPP with ongoing, long-term support through social prescribing and community engagement initiatives that increases attention to patients' wider social context and health choice architecture[8].", p. 12, ll. 7-17


1. Kleinsinger F. The Unmet Challenge of Medication Nonadherence. *TPJ*. 2018;22(3):18-033. doi:10.7812/TPP/18-033
2. Bullock JG, Green DP. The Failings of Conventional Mediation Analysis and a Design-Based Alternative. *Advances in Methods and Practices in Psychological Science*. 2021;4(4):251524592110472. doi:10.1177/25152459211047227
3. Miles LM, Hawkes RE, French DP. How the Behavior Change Content of a Nationally Implemented Digital Diabetes Prevention Program Is Understood and Used by Participants: Qualitative Study of Fidelity of Receipt and Enactment. *J Med Internet Res*. 2023;25:e41214. doi:10.2196/41214
4. Poupakis S, Kolotourou M, MacMillan HJ, Chadwick PM. Attendance, Weight Loss, and Participation in a Behavioural Diabetes Prevention Programme. *IntJ Behav Med*. Published online January 11, 2023. doi:10.1007/s12529-022-10146-x
5. Koutoukidis DA, Barron E, Stevens R, Aveyard P, Valabhji J, Jebb SA. Association between the month of starting a weight management program and weight change in people at high risk of type 2 diabetes: A prospective cohort study. *Obesity*. 2023;31(6):1707-1716. doi:10.1002/oby.23762
6. NHS Digital. Diabetes Prevention Programme: Non-Diabetic Hyperglycaemia, January to December 2021. *National Diabetes Audit* https://files.digital.nhs.uk/1E/852B7B/NDA_NDH_2021-22_Q3_01.01.21 to 31.12.21.xlsx (2022).
7. Whelan, M. & Bell, L. The English national health service diabetes prevention programme (NHS DPP): A scoping review of existing evidence. *Diabet. Med.* **39**, (2022).
8. Calderón-Larrañaga, S. *et al.* Unravelling the potential of social prescribing in individual-level type 2 diabetes prevention: a mixed-methods realist evaluation. *BMC Med.* **21**, 91 (2023).


## 2. *in addition to potential causal pathways and mediators, are there any time varying confounding that also need to be fleshed out? or represent threats to validity.*

Thank you for this important remark. A time-varying variable that may pose the risk of confounding is the increasing rate of screening through HbA1c testing. This is particularly relevant for the difference-in-differences analysis: supplementary figure S11 shows the increasing share of patients with a recorded HbA1c test (compared to the total patient population within a given practice) from April 2015 to March 2020 stratified by roll-out wave. Consequently, we performed a robustness analysis controlling for HbA1c testing rates in each practice. This analysis is described in the manuscript as follows:

"In line with changes in NICE guidelines[9], the share of patients who have a recorded HbA1c test in each practice increased from 2015 to 2020. The increasing trends were similar across waves, except for wave 2 practices in 2019/2020 that were lower compared to wave 3 practices (fig. S11). Hence, as a robustness analysis, we applied a conditional parallel trends assumption; that is, we assume only that practices with the same share of patients having taken a HbA1c test would follow the same trend in HbA1c in the absence of treatment. This resulted in similar effect estimates (Extended Data Table 6b)." p. 26, ll. 5-10

Related to time-varying confounding is also the circumstance that patients above the threshold are more likely to have their HbA1c tested. This is relevant to the validity of the exclusion restriction that needs to be met to arrive at causal effect estimates in regression discontinuity analyses. We discuss this and the fact our placebo analysis – a negative exposure control scenario – substantially mitigates this concern at two points in our main text:

- "First, we did a placebo analysis using an alternative patient cohort that is comparable in inclusion criteria to our primary analysis but who had baseline HbA1c measures in 2014 and 2015. During this period, patients were already identified with prediabetes based on the same threshold and public health guidance. However, the key difference was that there were limited intensive lifestyle programs in place to which GPs could refer their patients. This led to a scenario where key potential confounders such as discontinuities in outcome availability and number of consultations could already be observed (fig. S5), while the hypothesized causal mechanism—the NHS DPP—was 'deactivated' (i.e., a negative exposure control scenario; fig. S6a). We find that being eligible for the NHS DPP in these years in

which the NHS DPP was not yet available did not result in an improvement in HbA1c and BMI (fig. 6b-c)." Section 'Can we rule out confounding and bias?', pp. 7-8, ll. 27f.

- "Another potential limitation is violation of the exclusion restriction, whereby treatments or exposures other than intensive lifestyle counseling are affected by crossing the cutoff. While we could not precisely control for the relationship between drug dosage and secondary outcomes, we are confident that the observed health effects were not attributable to increased medication following treatment assignment as effect estimates were robust to adjusting for drug prescriptions. **Further, it is likely that patients classified as high risk for progression to T2DM have been under closer observation by GPs and that GPs may have taken additional steps to mitigate the risk. We empirically demonstrate that this is unlikely to be a substantial confounder in our analysis because we found no health improvements from crossing the prediabetes threshold prior to NHS DPP roll-out. Similarly, outcomes such as HbA1c or BMI were measured more frequently by GPs in the follow-up period if patients had crossed the threshold. However, testing a diverse set of so-called negative outcome controls—variables that we did not expect to be plausibly affected by referral to the NHS DPP—suggests no systematic confounding**." p. 12, ll. 18-28

***Referee #2*** *(Remarks to the Author):*
*-The paper addresses a significant topic and utilize innovative methods.*
*-The authors made several changes in response to the previous reviewers which strengthened the manuscript. Notably, they have clarified throughout the manuscript that the study evaluates effects of program referral rather than the program itself; they have included stratified analyses to examine potential effect modifiers; they have provided detailed information about DPP implementation; and they have examined effects at a variety of time points after referral.*

3. *A large amount of important information is contained in the supplemental files that cannot be properly discussed in the paper (and cannot fit in a single manuscript). For example, the authors make many general statements about results presented in the supplemental files (for the sake of brevity) that overlook important findings. For example, the authors state patients who referred and started the program are similar to patient who were referred and did not start the program yet table S15 shows some statistically significant differences between the groups. It is also difficult to keep track of primary and secondary cohorts, analyses, etc. in the supplementary files because of the large number of tables and figures; there is not enough space to call out and/or discuss each item in the primary manuscript.*

We share the reviewer's concern that some findings cannot be discussed in-depth for the sake of brevity. However, we would like to point out that statistical tests such as the Chi-Square test (that we used to compare patients who were referred to the NHS DPP and started the intervention with patients who were referred but did not start the intervention) are sensitive to sample size, implying that using a large enough sample (in our case a sample exceeding 20 000) also (proportionally) trivial differences will turn up as statistically significant. We, thus, present standardized mean differences (SMD, also known as Cohen's *d*) which is a standard effect size to quantify group differences. All SMD except for the average yearly number of GP consultations are below 0.1, which is why we state that "Relevantly, patients who started the treatment are largely similar to patients who were referred but did not start the treatment, in particular on variables that may indicate heightened personal motivation or health literacy such as attending preventative cancer screening services or vaccine uptake (table S4) […].", p. 7, ll. 5-8

Prompted by the reviewer's concern that is difficult to keep track of the many analyses, which is shared by the statistical reviewer, we now include an overview of the key results in the Extended Data Table 7. We have also re-organized the Supplementary Information including the presentation of the secondary cohort, which is now included as a separate section in the Supplementary Information (S3) to increase the ease in accessing and understanding the comprehensive results.

4. *The authors included a historical control (placebo analyses) to look at the influence of prediabetes diagnosis on clinical outcomes when DPP referral was not available; these analyses are very helpful in showing that prediabetes awareness alone does not lead to the observed outcomes. However, the authors also briefly mention rates of testing for prediabetes increased over time as providers were increasingly able to refer to the DPP (i.e., prediabetes testing increased with increased availability of the DPP for patients found to have prediabetes).* ***This finding deserves more discussion—could increased availability/access to DPP have effects on provider behaviors (screening and referral) and provider/patient perceptions of labeling/diagnosing a precondition during the evaluation period? Could these effects influence the study results?***

We thank the reviewer for their appreciation of the historical control ('placebo analysis'). We actually believe that this analysis does not only help to establish that prediabetes awareness alone does not lead to observed outcomes, but also that changes in provider behavior (i.e., increases in prediabetes testing as well as the number of consultations)—in particularly toward those above the threshold, which was also already the case during the 2014-2015 period, see supplementary figure S5—did not lead to health improvements. Nonetheless, we acknowledge that increased access to the DPP may have had effects on provider as well as patient perception and behavior that we have not already observed in our historical control, and that we, thus, cannot entirely preclude that such changes may have had an influence on outcomes. However, if it would be, for example, the case that a provider or patient is much more alert to their patient's or their own health, we would also expect to see effects on other health outcomes that are not plausibly improved by lifestyle changes. We demonstrate that this is not the case (Extended Data Table 5). We discuss this in the manuscript as follows:

"Further, it is likely that patients classified as high risk for progression to T2DM have been under closer observation by GPs and that GPs may have taken additional steps to mitigate the risk. We empirically demonstrate that this is unlikely to be a substantial confounder in our analysis because we found no health improvements from crossing the prediabetes threshold prior to NHS DPP roll-out. Similarly, outcomes such as HbA1c or BMI were measured more frequently by GPs in the follow-up period if patients had crossed the threshold. However, testing a diverse set of so-called negative outcome controls—variables that we did not expect to be plausibly affected by referral to the NHS DPP—suggests no systematic confounding. We further conditioned our results on variables that may function as a proxy for health literacy, health services utilization, and personal motivation, which did not substantially change our results, and performed a robustness analysis restricting our sample to patients with regular GP visits before their treatment assignment.", pp. 12-13, ll. 22f.

However, there is one exploratory result, which may correspond to what the reviewer has pointed out: When using recorded diabetes diagnoses as an outcome, referral to intensive lifestyle counseling actually leads to an *increase* in the probability of being diagnosed with T2DM (tables S2-3). We believe that is likely that a stronger focus on identifying people who are at risk of diabetes—and labelling them as such in combination with improved follow-up within the program—had led to an initial increase in the incidence of T2DM independent of health improvements. It has also been demonstrated that diagnoses in electronic health records may be less reliable than biochemical measures. We, thus, mainly rely on objective measures as our primary and secondary outcomes. Upon reflection and again, for the sake of brevity, we have decided to not discuss this in-depth in the Discussion. However, we have included the following in the Results section: "Analyses using T2DM, hypertension, and hyperlipidemia incidence as outcomes are presented in the Supplementary Tables S2-3 and show an increase in T2DM diagnoses at the threshold. We interpreted these results cautiously since diagnoses in electronic health records may be less reliable than biochemical measures[52] and a stronger focus on identifying people who are at risk of diabetes is likely to initially increase the incidence of T2DM independent of health improvements[53]." p. 9, ll. 4-8

Lastly, we discuss increasing HbA1c testing as a potential time-varying confounder in the difference-in-differences analysis. Specifically, supplementary figure S11 shows the increasing share of patients with a recorded HbA1c test (compared to the total patient population within a given practice) from April 2015 to March 2020 stratified by roll-out wave. Consequently, we performed a robustness analysis controlling for HbA1c testing rates in each practice. This analysis is described in the manuscript as follows:

"In line with changes in NICE guidelines[9], the share of patients who have a recorded HbA1c test in each practice increased from 2015 to 2020. The increasing trends were similar across waves, except for wave 2 practices in 2019/2020 that were lower compared to wave 3 practices (fig. S11). Hence, as a robustness analysis, we applied a conditional parallel trends assumption; that is, we assume only that practices with the same share of patients having taken a HbA1c test would follow the same trend in HbA1c in the absence of treatment. This resulted in similar effect estimates (Extended Data Table 6b)." p. 26, ll. 5-10

1. Persson R, Vasilakis-Scaramozza C, Hagberg KW, et al. CPRD Aurum database: Assessment of data quality and completeness of three important comorbidities. *Pharmacoepidemiology and Drug Safety*. 2020;29(11):1456-1464. doi:10.1002/pds.5135

***Referee #3** (Remarks to the Author):*

*This is a very thoughtful study, with elegant application of several strong quasi-experimental designs, providing internally valid suggestion that referral to diabetes prevention programmes results in improved HbA1c and risk factors but not complications.*

*The paper is well written, clearly presented, and brings some methodologic diversity to this question. Prior reviews have focused on addressing potential confounding and selection biases, for which these designs can never fully address, but nevertheless I consider the responses of the authors on these issues to be adequate. The additional questions and limitations in the paper are the following:*

5. *The background introduction is well written in general but ignores the large and growing literature of evaluations of the UK NDPP as well as other settings.*

It was not our intention to ignore the existing literature of evaluations of the English National Diabetes Prevention Programme. However, the evaluations largely consist of observational, single-arm pre-post studies that do not allow for causal interpretation. Given space limitations, we have now opted for the following:

- We include Supplementary Information 1.1, where we present existing studies (informed by Whelan & Bell's (2022) scoping review as well as from an ongoing literature search that we have conducted during our project).
- We point towards this information in the Introduction: "By exploiting this eligibility threshold that draws a sudden distinction between the prediabetes and normal glycemia range, we can take advantage of existing large-scale routine health data while still obtaining causal effect estimates that are not vulnerable to confounding and measurement error (see Supplementary Information 1.1 for an overview of existing correlational evidence for the NHS DPP)[10,36]." (p. 4, ll. 1-5)
- We also expanded the first paragraph in the Discussion comparing our results to existing evidence as follows: "While clinical trials have shown that intensive lifestyle counseling is efficacious in improving cardiovascular risk factors in controlled research settings[3,18,47–49], we presented evidence that these health benefits can be successfully translated to and scaled up in routine care. In our study using electronic health records from over two million patients, we found causal evidence that referral to the largest behavior change program for prediabetes globally led to improved glycemic control and reductions in BMI, weight, HDL, and triglycerides. Systematic reviews and meta-analyses of controlled clinical trials studying effects of lifestyle interventions (**including diabetes prevention programs in US settings**) on weight loss and blood pressure in adults with prediabetes found effect sizes comparable to those in our study[47,50–52]. While evidence on improvements in glycemic control from controlled trials appears more mixed[48,51,52], our study substantiates indications from **earlier correlational studies that suggest beneficial effects of NHS DPP participation on HbA1c and weight (see Supplementary Information 1.1)**." (p. 11, ll. 1-12)

## S1.1 Overview of peer-reviewed articles, abstracts and selected program statistics

| Author, year | Study design | Key results | Limitation |
|---|---|---|---|
| Bakhai & Hopwood, 2020[21] | Program statistics from presentation (single-arm, pre-post design) | HbA1c reduction of 2.0 mmol/mol and weight reduction of 3.3 kg for program completers (minimum of 60% of session attended) | No causality established |
| Koutoukidis et al., 2022[5] | Prospective cohort study (single-arm, pre-post design) | Weight reduction of 2 kg for participants who attended at least one session | No causality established |
| Marsden et al., 2022[22] | Prospective cohort study (single-arm, pre-post design) | HbA1c reduction of 2.1 mmol/mol and weight reduction of 3.6 kg for program completers (final session or final measurement plus minimum of 60% of session attended) | No causality established |
| McManus et al., 2022[23] | Difference-in-difference event study (staggered treatment design) | Incidence rates of T2DM in program-participating practices were significantly lower than would have been expected in the absence of the DPP (difference-in-differences Incident Rate Ratio (IRR) = 0.938 (95% CI 0.905 to 0.972 and 0.927 (95% CI 0.885 to 0.972)) | Some causality established, annual practice-level T2DM incidence rates as outcome only |
| Murray et al., 2019[24] | Pilot study of the digital version of the NHS DPP; prospective cohort study (single-arm, pre-post design) | HbA1c reduction of 1.6 mmol/mol and weight reduction of 4.0 kg at 6 months | Digital delivery only; no causality established |
| Poupakis et al., 2022[4] | Prospective cohort study (single-arm, pre-post design) | Every additional session attended is associated with a weight loss of 0.281 kg | No causality established |
| Ravindrarajah et al., 2023[25] | Matched cohort study | Probability of not converting to T2DM at 36 months since referral was 87.3% (95% CI: 86.5% to 88.2%) for referred to the program and 84.6% (95% CI: 83.9% to85.4%) for not referred to the program | No causality established, coded diagnosis as outcome only |
| Valabhji et al., 2020[26] | Prospective cohort study (single-arm, pre-post design) | HbA1c reduction of 1.26 and weight reduction of 2.3 for initial participants (attended at least one session); HbA1c reduction of 2.04 and weight reduction of 3.3 for completers (minimum of 8 of 13 sessions attended) | No causality established |
| Wise, 2018[27] | Program statistics cited (single-arm, pre-post design) | Weight reduction of 3.3 for all participants and 3.7 kg for participants excluding those with a normal BMI | No causality established |

T2DM = type 2 diabetes mellitus. This table is informed by Whelan, M., & Bell, L. (2022). The English national health service diabetes prevention programme (NHS DPP): A scoping review of existing evidence. *Diabetic Medicine*, *39*(7), e14855, as well as an informal search on PubMed using the search string 'diabetes prevention programme AND England'.

6. *The finding of the threshold is clear enough, but it is not clear what assumptions are necessary to conduct the "scaling" of effect size from the eligible-group effect to the referred group effect size. It is not clear why the referred group effect size cannot be calculated based on observed data independent of the eligible denominator. Doesn't this also have to make the assumption that the referred person has no selection effects differentiating them from the eligible but not referred? If there is unmeasured differences there, what would the effect on main findings be?*

In a regression discontinuity design, it is not necessary to assume "that the referred person has no selection effects differentiating them from the eligible but not referred". This is because the threshold rule is an exogenous instrument (i.e., randomly assigns individuals scoring *just* below and above the threshold) that we can use to obtain an unbiased causal effect estimate for referral. However, the reviewer is correct that the estimated causal effect size of referral in the regression discontinuity analysis only applies to those who are eligible and accept the referral because of the threshold rule. Thus, the (unobserved) effect is not necessarily the same if those who were eligible but not referred were also referred. This is an inherent feature of a (fuzzy) regression discontinuity analysis, and we acknowledge this as a limitation:

"Inherent to any regression discontinuity analysis is the limitation that we can only estimate the causal effect for those who initiate the treatment *because* they crossed the eligibility threshold. This effect may differ from the (unobserved) treatment effects for patients who would have been referred to intensive lifestyle counseling regardless of baseline HbA1c levels (the "always takers"), for example, because of clinical symptoms, or patients who would not have participated in any program or counseling even if eligible (the "never takers"). Additionally, effects may not be generalizable to those further away from the HbA1c threshold that defines prediabetes. This is, however, not a concern in this study given that the NHS DPP specifically targets people with prediabetes, which in turn is defined by a fairly narrow HbA1c range. In our analysis, we mainly relied on the complier average causal effect (CACE), estimating the effect for those who were actually referred to an intervention. Given that there was a large percentage of individuals presenting above the HbA1c threshold who were not referred to any lifestyle counseling, it is important to not mistake the observed CACE effects for the population health effect of the NHS DPP." (pp. 11-12, ll. 25f.)

We have also newly included the Extended Data Table 7, where we present an overview of key results for our primary outcome. In this table, we also include a brief explainer of how to interpret each estimated effect.

If the reviewer or editor feels like we should further expand on this, please let us know.

7. *As described, the main finding of benefit is derived from an observation that people in the 6.0-6.4 A1c range saw greater improvement than < 6%. Could the characteristics, meds, health status, changes in risk etc. of the group in the 6.0 to 6.4% category also be affected by changes in the category of them (i.e, diagnosis of diabetes at the 6.5% level)? In other words, if there were shifting diabetes diagnosis, particularly an increased in detection at lower levels than this would pull people from the upper end of the 6.0 to 6.4 range into the diagnosed category, could this result in a spurious decrease in the A1c of the pre-diabetes group?*

If we understood the reviewer's comment correctly, they are concerned about the fact that patients at the upper end of the prediabetes range (HbA1c: 6.0-6.4% or 42-47 mmol/mol) may be diagnosed earlier, which may lead to earlier treatment and in turn better health outcomes (e.g., HbA1c levels)—independent of the program's effects. The reviewer's reasoning seems to suggest that we compute effect based on the average effect in each range (< 6%, prediabetes range, > 6.5%). However, in the regression discontinuity, in particular, we estimate the effect at the threshold (6.0% or 42 mmol/mol), considering only patients in a close range ('bandwidth') around the prediabetes threshold and systematically vary this bandwidth, so that the effect is not sensitive to including patients at the upper end of the diabetes range (see Figure 2).

8. *If there is adequate data to assess change in A1c, risk factors, AND complications, its not clear why the paper would not generate an effect size for diabetes incidence. At the same*

*time, the analysis related to complications does not seem particularly useful. If we actually saw a significant effect there, within 1 year, we would be more apt to assume that is reverse causality anyway, with people without complications more likely to be referred, rather than vice versa.*

We fully agree with reviewer regarding the (exploratory) outcome complications. We are, thus, not surprised to not have seen a significant/meaningful effect on complications in the comparatively short follow-up period. Regarding an effect size for diabetes incidence, we actually do include an effect size for the probability of being diagnosed with T2DM: When using recorded diabetes diagnoses as an outcome, referral to intensive lifestyle counseling actually leads to an *increase* in the probability of being diagnosed with T2DM (tables S2-3). We believe that is likely that a stronger focus on identifying people who are at risk of diabetes—and labelling them as such in combination with improved follow-up within the program—had led to an initial increase in the incidence of T2DM independent of health improvements. It has also been demonstrated that diagnoses in electronic health records may be less reliable than biochemical measures[1]. We, thus, mainly rely on objective measures as our primary and secondary outcomes. Upon reflection and again, for the sake of brevity, we have decided to not discuss this in-depth in the Discussion. However, we have included the following in the Results section: "Analyses using T2DM, hypertension, and hyperlipidemia incidence as outcomes are presented in the Supplementary Tables S9-12 and show an increase in T2DM diagnoses at the threshold. We interpreted these results cautiously since diagnoses in electronic health records may be less reliable than biochemical measures[11] and a stronger focus on identifying people who are at risk of diabetes is likely to initially increase the incidence of T2DM independent of health improvements[53]." p. 9, ll. 4-8

1. Persson R, Vasilakis-Scaramozza C, Hagberg KW, et al. CPRD Aurum database: Assessment of data quality and completeness of three important comorbidities. *Pharmacoepidemiology and Drug Safety*. 2020;29(11):1456-1464. doi:10.1002/pds.5135

***Referee #4**:*
*As per request from editor I have reviewed the paper from a statistical/causal inference perspective.*

** Main remark: I find the statistical analyses included very comprehensive and state-of-the-art. I would like to congratulate the authors on an impressive work.*

** Correctness: As far as I can tell the method are correctly applied. They do indeed build an different causal assumptions, which is a strength. However, it does also imply that the methods are not targeting exactly the same parameter. This imply that - in the extreme - we would not require they arrive at the same estimate. Weaker properties such as "same sign and significance" is perhaps all we can demand.*

** Use of the multiple methods: The many different methods employed has the natural consequence that the amount of model output is LARGE. This is unavoidable.* ***However, I think the authors could still make one more table or figure (yes I see the irony) which illustrated the estimates on the key outcome obtained from the different methods****. Right now it took me ages to find out exactly which estimates should be compared with which. This would also make it easier for the reader to assess if the agree, that the conclusions are indeed supportive.*

We thank the reviewer for their positive assessment of our statistical analyses. We think providing an illustration of the key results is an excellent suggestion and we have now included an overview table in our manuscript (Extended Data Table 7) where we show effects for our primary outcome from the different analyses, including a brief explanation of how to interpret each estimand.

**Reviewer Reports on the Second Revision:**

Referees' comments:

Referee #1 (Remarks to the Author):

The authors addressed my comments satisfactorily.

Referee #3 (Remarks to the Author):

I continued to think feel that this is an elegant, well conducted study, and arguably one of the most interesting quasi-experiments ever devoted to the area of chronic disease prevention.

However, I have one remaining concern, that was actually amplified by this most recent round. The decision to present effects of HbA1c and bury findings related to diabetes incidence in one sentence of the results and a supplemental table doesnt make sense. The authors cite reliability concerns for diabetes diagnosis and cite one study to support this. (Of note, the authors include BMI as a primary outcome here despite the dubious nature of its measurement throughout primary care, ie.,with clothing, keys, wallets, and all other sorts of combinations.). More importantly, though, are HbA1c values among the new cases part of both baseline and follow-up? If intervention causes HbA1c to go down and cases to go up, is it possible that the identification of new cases is removing the high HbA1c levels from the numerator at follow-up, accounting (at least in part) for the reduction in incidence. Whatever the case, with the availaiblity of diagnosies, medications, and laboratory data, the authors have all that his needed to assess diabetes incidence and report it as a main outcome.

Referee #4 (Remarks to the Author):

Thanks for a great extra table on comparing the estimates. It worked even better than I had hoped. And as expected there were no "surprises".

In short: all my concerns have been met.

## Author Rebuttals to Second Revision:

**Referees' comments:**

***Referee #1:***
*The authors addressed my comments satisfactorily.*

***Referee #4:***
*Thanks for a great extra table on comparing the estimates. It worked even better than I had hoped. And as expected there were no "surprises".*

*In short: all my concerns have been met.*

***Referee #3:***
*I continued to think feel that this is an elegant, well conducted study, and arguably one of the most interesting quasi-experiments ever devoted to the area of chronic disease prevention.*

*However, I have one remaining concern, that was actually amplified by this most recent round. The decision to present effects of HbA1c and bury findings related to diabetes incidence in one sentence of the results and a supplemental table doesnt make sense. The authors cite reliability concerns for diabetes diagnosis and cite one study to support this. (Of note, the authors include BMI as a primary outcome here despite the dubious nature of its measurement throughout primary care, ie.,with clothing, keys, wallets, and all other sorts of combinations.). More importantly, though, are HbA1c values among the new cases part of both baseline and follow-up? If intervention causes HbA1c to go down and cases to go up, is it possible that the identification of new cases is removing the high HbA1c levels from the numerator at follow-up, accounting (at least in part) for the reduction in incidence. Whatever the case, with the availaiblity of diagnosies, medications, and laboratory data, the authors have all that his needed to assess diabetes incidence and report it as a main outcome.*

**Our response:**
We thank all reviewers for their positive assessment of our study. We sincerely appreciate the reviewers' time and expertise in evaluating our work.

We identified two remaining concerns in the reviewer #3's comment.

First, we are happy to include diabetes incidence as a main outcome. It was not our intention to bury this findings as we consider a short- to mid-term increase in diabetes diagnoses a plausible phenomenon given intensified monitoring. The diabetes diagnosis outcome data are now included in Extended Data Table 3 and 4.

We have also added the following text to our Results section:

"Diabetes complications, emergency hospitalization for major adverse cardiovascular events, and mortality were not significantly reduced by being referred to intensive lifestyle counseling in exploratory analyses (Extended Data Table 3). The low incidence of adverse downstream events during our relatively short follow-up period, with 26 567 (1.3%) of patients dying and 36 567 (1.8% of those 2 037 384 linkable to HES data) having an emergency hospitalization for a major adverse cardiovascular event, resulted in low statistical power for detecting small, short-term changes in these outcomes. **We observed a significant increase in T2DM diagnoses at the threshold (RD = 0.39, 95% CI 0.21, 0.57; Extended Data Table 3). We interpreted these results cautiously since diagnoses in electronic health records may be less reliable than biochemical measures[37] and a stronger focus on identifying and monitoring people who are at risk of T2DM is likely to initially increase the incidence of T2DM independent of health improvements. Hypertension and hyperlipidemia incidence did not significantly change at the threshold (Supplementary Tables S2-3)**."
Page 8, lines 27 ff.

In addition, we added the following sentence to our Discussion sentence:

"This study has several other limitations. First and foremost, we observed considerable missingness in outcomes such as HbA1c or BMI, which were also measured more frequently by GPs in the follow-up period if patients had crossed the prediabetes threshold. This differential missingness in our outcome variables will be a source of

bias if individuals with missing outcome data have systematically different outcome values than those who do not. Having said that, we show that crossing the prediabetes threshold and subsequent closer monitoring of biomarkers by GPs was not associated with improvements in our outcome variables prior to NHS DPP roll-out. **Closer monitoring of patients, however, may be the reason for which we find that referral to intensive lifestyle counseling increased the probability of being diagnosed with T2DM independently of improvements in biomarkers.**" Page 12, lines 18 ff.

Second, we would like to clarify that **we do not exclude** new cases from follow-up. We agree with the reviewer that excluding new cases with possibly high HbA1c levels from endline measurements would introduce bias. Hence, in our endline HbA1c measurement all patients with an available record are included.