TECHNICAL APPENDIX

This appendix concerns the statistical analysis of the of the REaCH trials. Data from each trial were analysed separately. The document describes the analysis of the primary outcomes and summarises the statistical analysis plan; analysis of the secondary outcomes followed the same method with appropriate adjustment for the type of outcome.

Endpoints

The primary outcomes of the trial are:

1. Consultation rate by remote consultation (consultations per patient-month)
2. Consultation rate by face-to-face consultation (consultations per patient-month)
3. Prescribing rate (prescriptions per patient-month)
4. Trust in healthcare provider (PHBQ).

Estimands

We aimed to estimate the following treatment effects among the group of patients represented by the inclusion criteria:

1. The rate ratio for the rate of remote consultations of patients enrolled in clinics that have received the REaCH training package versus patients in clinics providing standard care.
2. The rate ratio for the rate of face-to-face consultations of patients enrolled in clinics that have received the REaCH training package versus patients in clinics providing standard care.
3. The rate ratio for the rate of prescriptions for patients enrolled in clinics that have received the REaCH training package versus patients in clinics providing standard care.
4. The mean difference in standardised PHBQ score between patients enrolled in clinics that have received the REaCH training package versus patients in clinics providing standard care.

Our treatment effects are all intention to treat and participants are assigned the treatment status of the cluster providing treatment at the time of the consultation.

Statistical analysis

The primary outcomes are of two types: count data (the consultation and prescription rates) and a continuous score (the patient trust outcome). For the former we specify the following log-linear model, for patient $i = 1, \dots, N$ in cluster $j = 1, \dots, J$ at time period $t = 1, \dots T$ with number of events $y_{ijt}^{(1)}$:

$$y_{ijt}^{(1)} \sim Poisson(\lambda_{ijt})$$

$$\lambda_{ijt} = \exp\left(\eta_{ijt}^{(1)}\right) \tag{1}$$

where η_{ijt} is the linear predictor. For the continuous outcome $y_{ijt}^{(2)}$ we specify:

$$y_{ijt}^{(2)} \sim N\left(\eta_{ijt}^{(2)}, \sigma^2\right) \tag{2}$$

The specification of the linear predictor in both models takes the form:

$$\eta_{ijt}^{(p)} = x_{ijt}'\beta + \delta D_{jt} + \alpha_j + \phi_i + \psi_{jt} \tag{3}$$

where x_{ijt} is a vector patient-level covariates, D_{jt} is an indicator equal to one if cluster j has the intervention at time t and zero otherwise, and $\alpha_j \sim N(0, \sigma_\alpha^2)$, $\phi_i \sim N\left(0, \sigma_\phi^2\right)$ and $\psi_{jt} \sim N\left(0, \sigma_\psi^2\right)$ are cluster, individual, and cluster-time and random effects respectively. The parameter δ provides an estimator for the treatment effects: $\exp(\delta)$ is the rate ratio.

Estimation

Algorithms for estimating the parameters of generalised linear mixed models can often fail. To ensure reliable estimates we will use two alternative fitting algorithms and compare results. We will use the R package lme4, which uses a penalised quasi-likelihood approach, and glmmrBase, which provides Markov Chain Monte Carlo Maximum Likelihood (MCMCML) methods. In the event of any disagreement or failure of lme4, we will use the MCMCML results, otherwise we will take the lme4 results.

Inference

We will report point estimates, confidence intervals, and p-values but not make any claims of "statistical significance" given recent strong arguments against doing so[2]. Our interpretation of the results will be based on the patterns of evidence in the context of the implementation and qualitative analyses.

P-values will be based on the null hypotheses $H_0: \delta = 0$ versus the two-sided alternatives $H_1: \delta \neq 0$ in each of the models defined in Equations (1)-(3). P-values and related statistics generally do not have the nominal type I error rates when the number of cluster is small (e.g.[4,5]). We therefore estimate p-values using a permutation test based approach. Given there are multiple primary outcomes we will also report adjusted p-values for multiple testing using a stepdown method, which provides an efficient means of controlling the family-wise error rate[6]. The full details of permutation-based inference, including confidence intervals/sets and corrections for multiple testing are given in Watson et al (2023)[7]. Here, we give a brief outline.

Our test statistic is a sum of residuals: $T = \sum_j \sum_t \sum_i (D^*_{jt}(y_{ijt} - \mu_{ijt}))$ where D^*_{jt} is the modified treatment indicator equal to one if the cluster had the intervention at time t and -1 otherwise. The parameter $\mu_{ijt} = h(\eta_{ijt})$ where h is the link function. To implement a permutation test, we re-randomise the clusters 10,000 and re-calculate the value of the test statistic under the null hypothesis, with all other "nuisance" parameters fixed to their maximum likelihood estimates. The p-value is derived by comparing the actual test statistic against the sample of permuted test statistics, as shown below. Confidence intervals can be derived from this method using the search procedure proposed by Garthwaite[8,9].

To describe extending the permutation-test based method to correct for multiple testing, let there be P hypotheses to be tested $H_1, \dots, H_P$ with associated test statistics T_p. We let the ordered test statistics be

$$T_{[1]} \geq T_{[2]} \geq \cdots \geq T_{[P]}$$

corresponding to hypotheses $H_{[1]}, \dots, H_{[P]}$. The family of hypotheses being tested is $K \subset \{1, \dots, P\}$. We first describe a stepdown procedure for the decision to accept/reject given a value for the type I error rate α as this provides a way of deriving the associated p-value that we will report. The stepdown method works by firstly testing if the joint null hypothesis that all null hypotheses are true by comparing the largest test statistic to some critical value $c_K\left(1 - \frac{\alpha}{2}\right)$. If it is smaller than this critical value we accept all null hypotheses otherwise we reject $H_{[1]}$ and test the remaining hypotheses as a new family in the same way. More specifically, the algorithm is

1. Let $K_1 = \{1, \dots, P\}$. If $T_{[1]} \leq c_{K_1}\left(1 - \frac{\alpha}{2}\right)$ then accept all hypotheses and stop, otherwise reject $H_{[1]}$ and continue.
2. Let K_2 be the indices of hypotheses not previously rejected. If $T_{[2]} \leq c_{K_2}\left(1 - \frac{\alpha}{2}\right)$ the accept all remaining hypotheses, otherwise reject $H_{[2]}$ and continue.
3. …

The critical values here are the $1 - \frac{\alpha}{2}$ quantiles of the distribution of the largest test statistic for the relevant family of hypotheses. Exact distributions may not exist, however we can derive them using a permutation testing approach. As before, we can use this distribution to determine the p-value.

At each stage of the stepdown algorithm we are testing the largest test statistic, that is $T_K = \max_{q \in K} T_q$ and then if the associated hypothesis is rejected, we create a new family of hypotheses that excludes the rejected hypothesis. If we re-randomise the clusters the for each permutation $m = 1, \dots, M$ we can re-calculate the desired test statistic, $T_K^{(m)}$, to create our reference distribution. The p-value is then:

$$p_{[k]} = \frac{1}{M+1}\sum_{m=1}^{M}\left(1 + I\left(T_K^{(m)} \geq T_K\right)\right)$$

That is to say that the reference distribution for the largest test statistic is the largest test statistics from the permutations, and so on.

In the main report we provide p-values and confidence sets both with and without multiple testing corrections.

Missing data

We had pre-specified analyses to evaluate the effect of any missing data. However, only 0.8% of observations were missing across our primary outcomes, and so no missing data analyses were performed as any effects would be negligible.

References

1. Hemming K, Taljaard M, Forbes A. Analysis of cluster randomised stepped wedge trials with repeated cross-sectional samples. Trials. 2017 Dec 1;18(1):101.
2. McShane, Blakeley B., et al. "Abandon statistical significance." *The American Statistician* 73.sup1 (2019): 235-245
3. Romano JP, Wolf M. Exact and approximate stepdown methods for multiple hypothesis testing. Journal of the American Statistical Association. 2005 Mar 1;100(469):94-108
4. Watson SI, Akinyemi J, Hemming K. Permutation-based multiple testing corrections for p-values and confidence intervals for cluster randomised trials. Statistics in Medicine. 2023 [In press]