Supplementary Appendix

Summary of RCT data collection and intervention design

The study at hand was conducted using data collected as part of a prior RCT. The details of the trial protocol were previously published, along with an estimate of the average intervention effects.[17] The Supplementary Information to that publication (available at www.nejm.org/doi/suppl/10.1056/NEJMc1806550/suppl_file/nejmc1806550_appendix.pdf) includes a detailed description of the study setting and population (Sections A and B). Similarly, Section C of that document includes a description of the randomization protocol, enrollment, and staffing for the trial implementation.

That document (Section D) also describes the intervention provided to patients enrolled in the treatment arm. In summary, each day at a time specified by the patient (in consultation with their doctor) they received a message reminding them to take their TB medication and to subsequently verify their adherence using the platform's USSD code. If the individual failed to verify, they were sent a second reminder an hour later. In case of further absence of verification, they received a third reminder an hour later. If they failed to verify yet again, they were marked as non-adherent. Subsequently, support sponsors messaged those who had been non-adherent for less than 24 hours and called those who had been non-adherent for more than 24 hours. Finally, the same section describes the technical implementation of the patient-platform interface, along with justification for those design choices.

For this intervention (and most others DAT-based ones) a mobile phone is needed. However, a key feature of the intervention is that it works on any feature phone (so patients do not need access to a smart phone). Note that USSD stands for *Unstructured Supplementary Service Data*, often referred to as "feature codes". USSD is a communications protocol used by cellular telephones, relying only on a GSM connection. Therefore, no data bundles (or access to internet) are needed. USSD platforms enhance security of health and other private data because no data is stored on the individual's device. Furthermore, USSD platforms obviate the need for the individuals purchase new hardware, or to install and maintain special software, and thus reduce barriers to adoption. Furthermore, we note that the use of USSD is ubiquitous in low- and middle-income countries including Kenya, where it is used for banking, for paying bills

(including tuition, utilities, etc.), as well as to interact with official programs such as the National Hospital Insurance Fund or the National Social Security Fund. One reason for the popularity of this communication standard is the prevalence of mobile phones in Kenya, which is over 95%, as we mention in the *Study Setting* section of the paper.

The analysis of the trial outcomes (described in Sections E and F of the SI to the original study) compares the individuals enrolled in the intervention with individuals who received the standard of care, using the same outcomes as we use in the present study. The trial found that patients enrolled on the platform had a rate of unsuccessful outcomes of 4.22% (24 patients out of 569 enrolled, with a standard error of 0.84%). In the control arm, the rate of unsuccessful outcomes was 13.08% (70 patients out of 535, with a standard error of 1.46%). The difference between the groups was statistically significant ($p<0.001$). Finally, the aforementioned Supplementary Information document includes additional results, e.g., analyses of the intervention effect for bacteriologically confirmed patients, focusing just on loss to follow up as an outcome, and using a multivariate regression specification.

Boutilier JJ, *et al. BMJ Global Health* 2022; 7:e010512. doi: 10.1136/bmjgh-2022-010512

Model details

The causal forest was fit using the grf package in R. We used 5000 trees and tuned all other model parameters ("sample.fraction", "mtry", "min.node.size", "honesty.fraction", "honesty.prune.leaves", "alpha", "imbalance.penalty") using internal cross-validation ("tune.parameters='all'"). We used the following tuning parameters: 2000 trees in each mini-forest ("tune.num.trees"), 100 forests ("tune.num.reps"), and 5000 random repetitions ("tune.num.draws"). Please see Athey and Wager (2019), or the grf reference documentation for more details.[18]

The causal forest makes one primary assumption about unconfoundedness and Figure A1 shows that the model satisfies that assumption. We assessed the goodness-of-fit using the statistical test proposed by Athey and Wager (2019).[18] The null hypothesis of this test is that the mean forest prediction is incorrect. We find a test statistic of 0.972 with a standard error of 0.210 corresponding to a t-statistic of 4.635 and a p-value of less than 0.001. The test suggests the model is well-fit since we reject the null hypothesis.

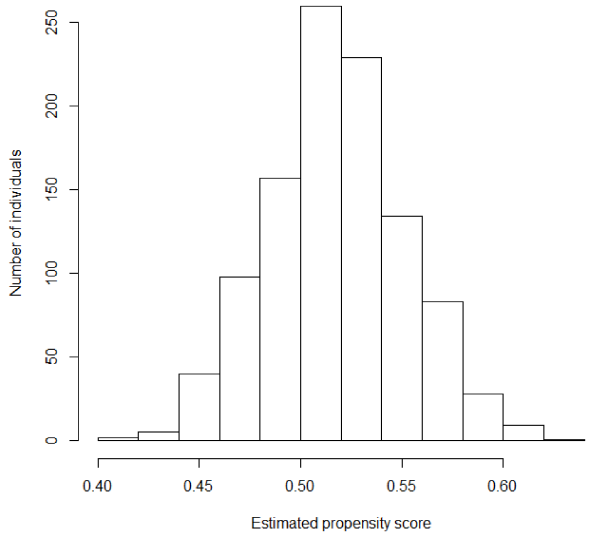

Figure A1. Individual propensity scores estimated by the causal random forest.