decomposition and the estimation of effects of covariates at both levels. However, random intercept models require exogeneity of exposure as an identifying assumption, and we have substantial background knowledge to suggest that this assumption would be violated in this case, because individuals participate in Juntos for reasons that are not reflected in measured covariates.

The aim of matching is to achieve 'conditional exchangeability', which was manifest as balance, as indicated by a lower standardized mean difference in measured covariates between the treatment and control groups. We assessed several matching algorithms, including matching with or without replacement, matching each exposed observation to one or more than one control and matching with or without a caliper, and we also allowed for transformations of and interactions between covariates. Matching each treated observation to control observations within a 10% caliper of the estimated propensity score with replacement provided the best balance of covariates. The matching ratio (whether 1:1 or some other ratio, 1:M) simply reflects the best balance achieved for the target population for the causal question. For example, if one wants to answer the causal question about the exposed population in relation to the counterfactual that these same individuals had not been exposed, then one should indeed use all exposed individuals, matched to one or more unexposed observations. The advantage of 1:M over 1:1 is simply the use of more of the unexposed observations, and therefore some improvement in precision, but it does not affect the validity of the results.

For the construction of the propensity score, we included variables that were related to the outcome of interest, but excluded variables that were a consequence of the exposure. When selecting variables for propensity score analyses, it is recommended to include confounders, specifically characteristics that are common causes of the exposure and outcome. Additionally, including variables unassociated with the exposure has been shown to increase the precision of estimates if they predict the outcome (the same logic applies in a randomized trial). Variables that are a consequence of the exposure (e.g. mediators) should never be included because they could induce bias because of collider stratification or result in an underestimate of the total effect (Brookhart *et al*. 2006). We did not include district as a matching covariate as the sample size for each district was small.

We estimated the effect of Juntos on maternal and child health in the matched subsets on the prevalence ratio scale by regressing each outcome on the treatment using generalized linear models (GLM). For the district-level analyses, these models were fitted with robust variance to account for the clustering of observations within districts (Williams 2000). Additionally, because matching with replacement allows for some observations to enter the analysis more than once, these analyses frequency weighted control observations by the number of times they were selected as a match (Dehejia & Wahba 1998). To evaluate the differences between Juntos and non-Juntos districts in the prevalence of outcomes prior to the implementation of Juntos, we compared prevalence proportions from 2007 (pre-implementation) and 2013 (post implementation).

All statistical analyses were conducted using Stata version 12.1. Propensity score methods were applied using Stata's -psmatch2- command (Nichols 2007).

### Sensitivity analyses

We conducted sensitivity analyses to test the robustness of our main findings in the district-level analysis. The first analysis was restricted to the participants that lived in a district with information on our outcomes before implementation in 2007. Because the prevalence of chronic malnutrition was a criterion used to select a district for receipt of the Juntos programme, we added the prevalence of chronic malnutrition of children in the district before the implementation of the programme to the estimation of the propensity score. In the second analysis, we conducted a propensity score matched analysis for each outcome using the prevalence of the outcome in each district before the implementation of the programme (i.e. for women, the prevalence of anaemia, underweight and overweight, and for children, the prevalence of acute malnutrition, anaemia and complications after delivery in year 2000).

We did not perform any adjustments or imputations for missing data because most (97.5%) missing values were for outcome variables (anaemia, underweight,