specific question in the DHS questionnaire. For the district-level analysis, the treatment variable was living or not in a district where Juntos was already implemented in the year the survey was conducted.

Outcomes

The DHS collects information about maternal and child health. For mothers, outcomes included anaemia and measured height and weight. Height and weight were used to calculate body mass index (BMI) and to classify respondents as underweight (BMI $\leq$ 18.5) or overweight (BMI $\geq$ 25) following the World Health Organization (WHO) International Classification system (Guilbert 2003). For children, outcomes included the incidence of complications after delivery, anaemia on women and children and acute malnutrition, defined as having a measured weight-for-height less than two standard deviations from the mean for normal children based on WHO growth standards (WHZ $<$ -2) (Tazza & Bullón 2006). Haemoglobin levels were measured by DHS with the HemoCue system. This is a simple and reliable test that uses photometric detection. Haemoglobin levels were then adjusted by altitude of residence. Anaemia was defined as adjusted haemoglobin levels below 11 g/dL. Trained personnel measured haemoglobin in participants, and height and weight in children (Instituto Nacional de Estadística e Informática 2015).

To evaluate compliance with conditions for staying in the programme, we included a variable for being born and having checkups at a health centre. In addition, we included compliance with current vaccination requirements (BCG, DPT, polio and measles).

Covariates

We accounted for potential confounding by calculating a propensity score based on maternal, child and household-level characteristics. Maternal characteristics included age at interview, height, educational attainment, literacy and reproductive characteristics, including the total number of children born and giving birth to more than two children in the past 5 years. Child characteristics included age at interview and height and weight at birth. Household characteristics included rural vs. urban residence, number of household members, household poverty and experiencing a child death in the family. We also controlled for year of interview, categorized as 2009–2010 compared to 2011–2012 in the individual analysis; and 2007–2009 vs. 2010–2013 in the district level analysis. The characteristics of the Juntos programme did not change significantly between these years. Categorizing year of interview dichotomously produced better matching in the propensity score. We did not include time of enrollment in JUNTOS because it was collinear with the variable 'year of interview'.

Statistical analyses

We used propensity score matching to (i) achieve balance in the distributions of measured covariates between the treatment and control groups and (ii) avoid extrapolation by limiting inference to regions of 'common support'. This involves an iterative process that begins with the estimation of the propensity score. For the individual-level analyses, the propensity score was defined as the predicted probability of enrollment in Juntos, estimated separately for mothers and their children, as a function of the measured maternal, child and household-level characteristics defined above. For the district-level analyses, the propensity score was defined as the predicted probability of living in a Juntos district, estimated separately for mothers and children, conditional on the same measured covariates.

The main advantages of using propensity score matching are the opportunity for non-parametric contrasts and flexible modelling of potential confounding in the first stage of the propensity score model. Another distinct advantage is the allowance for balance checks. It is true that the analytic sample tends to be reduced to the matched observations, but this is not necessarily a weakness. Indeed, for heterogeneous effect estimates, this helps minimize bias in the estimate of a specific target-population effect estimate. One may pay a price for this improved validity in the form of reduced precision, but in our large data set, it is arguably better to aim for a more unbiased estimate, rather than a more precise one.

We estimated propensity scores using multivariable logistic regression models and then matched on the propensity score. A multilevel analysis permits variance