**Supplemental Digital Content 1: Propensity Score Matching Approach**

Employer-level selection bias was an important consideration for this study. To reduce this potential bias, we included baseline costs, a proxy for use, in the propensity score matching procedure. Consequently, in our matched sample, high-utilizing HDHP members are matched to similar high-utilizing HMO members. To further investigate this potential threat to validity, we examined baseline trends in utilization for the main outcomes, below, and found the study groups followed similar baseline trajectories in use, indicating the central underlying assumption of the difference-in-differences framework was met.

Our propensity score model included the following variables: age, sex, family or individual plan status, health status (Adjusted Clinical Groups score), employer size, neighborhood socioeconomic characteristics, baseline copay levels, total member health plan costs at baseline, and secular changes. We performed an exact match on the characteristic of having one of four chronic diseases to ensure adequate representation of these groups. We also performed an exact match on the variable of association plan status because these very small employers purchase health plans in a different manner (through small business associations rather than directly from Harvard Pilgrim) and we have slightly less information on enrollees (such as whether they had access to Health Reimbursement Arrangements).

We used propensity score matching to generate a similar comparison group, although we found that residual baseline differences between the cohorts remained (Table 1). To investigate whether controlling for residual differences generated more valid effect estimates compared to unadjusted estimates, we ran adjusted and unadjusted models using Poisson regression. (This approach of using propensity score balancing followed by regression adjustment has been promoted by Robins and collaborators as protecting against residual baseline differences due to imperfections in the propensity score model.) However, we found that results were nearly identical and that interpretation did not differ. We therefore present the unadjusted results in the paper (Table 2).





















**References**

Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005 Dec;61(4):962-73.