SUPPLEMENTARY DATA

I. Study Group Construction and Deductible Imputation Algorithm

To determine employer deductible levels, we used a benefits type variable that we had for most smaller employers (with approximately 100 or fewer employees). For larger employers, we took advantage of the fact that health insurance claims data are the most accurate source for assessing out-of-pocket obligations among patients who utilize health services. Our claims data contained an in-network/out-of-network deductible payment field. For patients who use expensive or frequent services, the sum of their yearly deductible payments add up to clearly identifiable exact amounts such as $500.00, $1000.00, $2000.00, etc. When even several members have these same amounts, it provides strong evidence that the employer offered such an annual deductible level. It is also possible to detect employers that offer choices of deductible levels when multiple employees have deductibles at two or more levels, such as 20 employees with an exact annual amount of $1000.00 and 12 employees with $500.00. For employers with at least 10 enrollees, we therefore summed each member's in-network deductible payments and number of claims over the enrollment year and assessed other key characteristics such as percentage with Health Savings Accounts. We randomly selected half of the employer data set that contained both our calculated employer characteristics (independent variables, below) and actual annual deductible levels from the benefits table (dependent variable, after categorization; below). We then used a logistic model that predicted the 3-level outcome of deductible <=$500/$500-$999/>$1000 (again, dependent variable) based on multiple aggregate employer characteristics (independent variables) such as the first and second most common whole number deductible value, the percentage with Health Savings Accounts or Health Reimbursement Arrangements, the median deductible payment, the percentage of employees using services, the employer size, the percentage of employees with summed annual deductible amounts (from claims data) between $100 to ≤$500/ >$500 to <$1000/ ≥$1000 to ≤$2500/ >$2500, etc. This predictive model output the probability that employers had deductibles in the three categories (summing to 1) and we assigned the employer to the level that had the highest probability. If we detected employers that had enrollees with whole number deductible levels both above and below $500 (e.g. $250.00 and $1500.00), we assigned the employers' category as "choice." If 100% of employees had Health Savings Accounts, we also overwrote any previous assignment to classify the employer as a high-deductible employer. We tested the predictive model on the other half of the sample for which we had actual deductible levels from the benefits table (Table 1). At employers with 75-100 enrollees, we found sensitivity and a specificity of over 96%. The sensitivity and specificity would be expected to be even higher at employers with more than 100 enrollees (because more claims data would be available to provide evidence of deductible levels), but we were unable to test this because the dataset for which we had actual deductibles included employers with generally 100 or fewer enrollees.

Supplementary Table 1. Validation of deductible imputation algorithm.

	Gold Standard[a]=high-deductible (n)	**Gold Standard=low-deductible (n)**
We imputed high-deductible	611,541	14,335
We imputed low-deductible	24,017	465,120
	High-deductible	**Low-deductible**
Sensitivity	96.2%	97.0%
Specificity	97.0%	96.2%

SUPPLEMENTARY DATA

Positive Predictive Value	97.7%	95.1%

[a]Gold standard was a benefits variable specific to each employer derived from a benefits table and obtained from the health insurer via the data vendor.

SUPPLEMENTARY DATA

Rationale for low- and high-deductible cutoff values: when health savings account-eligible HDHPs came to market in 2005-2006, the Internal Revenue Service set the minimum deductible level for qualifying HDHPs at $1050 (which could be adjusted upward for inflation annually). The range of this minimum deductible during our study period was $1050-$1200. For these reasons, we defined HDHPs as annual individual deductibles of at least $1000 (otherwise health savings account plans would be excluded). In addition, choosing this cutoff (as opposed to e.g. $2000) also improves the sensitivity and specificity of the imputation because this is common deductible level and more enrollees per employer meet this threshold. This cutoff is also a "real-world" deductible *minimum* that allows the most generalizable results. We did not create a separate imputation algorithm for deductible levels of e.g. >=$2000 due to concerns that a less sensitive and specific algorithm would lead to biased effect estimates and a smaller HDHP sample size. It is important to note that $1000 was the ***minimum*** annual deductible level and not the mean deductible level. We cannot calculate the mean deductible level of the HDHP group directly but would expect it to be in the range of approximately $1500 to $2000. We defined traditional plans as having deductible levels of ≤$500 after determining that a threshold of ≤$250 would lead to an inadequate sample size for the control group. Again, the mean deductible level of the control group members would be lower than $500.

After preferentially assigning actual deductible levels (from our small employer benefits file) then imputed deductible levels at the employer plan year level, we began with 1,830,665 employer plan years. We excluded 201,230 plan years (11%) that included deductible levels other than only low or only high. Among the remaining 1,629,435 plan years, we excluded 191,519 (12%) that did not have 2 years of continuous enrollment. Finally, from the remaining 1,437,916 employer plan years, we excluded 549,638 (38%) that were not transitions of low deductible to low deductible or low deductible to high deductible. Most of these exclusions were due to employers having high deductibles at their initial appearance in our dataset and remaining in high deductible plans.

Our HDHP group therefore comprised the enrollment years of employers that had a year-on-year transition from low- to high-deductible coverage (from $500 or less to $1000 or more). Some employers had multiple eligible index dates (e.g., multiple low-to-low deductible years or both low-to-low and low-to-high deductible years). In these cases, we randomly assigned employers to the HDHP or control pool then randomly selected one of their index dates (and their corresponding before-after enrollment years).

We identified patients with diabetes age 12 to 64 as defined by detection of 1 inpatient or 2 outpatient diagnosis codes for diabetes (Table 2), or the dispensing of insulin or at least one oral hypoglycemic medication other than metformin alone, between 6 months before to 6 months after the beginning of members' baseline period.

Supplementary Table 2. Diagnostic and medication codes and used to define the diabetes cohort

Code Description	ICD-9-CM, DRG, or AHFS Code
To include in denominator:	
Diabetes diagnosis	250.0-250.93
Polyneuropathy in diabetes	357.2
Diabetic retinopathy	362.0
Diabetic cataract	366.41
Diabetes mellitus complicating pregnancy	648.03, 648.04
Uncomplicated diabetes, age over 35[1]	294
Uncomplicated diabetes, age 35 and under[1]	295
Diabetes with MCC[2]	637
Diabetes with CC[2]	638
Diabetes without CC/MCC[2]	639

SUPPLEMENTARY DATA

To exclude from denominator:	
Polycystic ovary syndrome	256.4
Other specified disorders of pancreatic internal secretion	251.8
Poisoning by adrenal cortical steroid	962.0

[1]Used before 10/1/2007; [2]Used on or after 10/1/2007
Abbreviations: ICD-9-CM, International Classification of Diseases, 9th Revision, Clinical Modification; DRG, Diagnosis Related Group; AHFS, American Hospital Formulary Service

II. Coarsened Exact Matching Approach

Coarsened exact matching helps control for the confounding influence of baseline study group differences by reducing imbalance on matching variables between the intervention (e.g., HDHP) and control groups.[1,2] We used a weighted coarsened exact match[1,2] on employer- and member-level propensity[11,12] to join HDHPs, baseline out-of-pocket expenditures, and members' baseline high- and low-severity emergency department visit and cost trends. We included all variables in the match as continuous variables. The logistic model for calculating employer propensity[6-9] to join a HDHP predicted this likelihood based on calendar index month, employer size (<50, 50-99, 100-249, 250-499, 500+); percentage of: women, members in income strata, education strata, age strata, race strata, and region strata; employer baseline cost level and trend; average employer Adjusted Clinical Groups (ACG) score; and outpatient copay. We constructed the corresponding member-level propensity model to ensure contemporaneous study groups as well as to balance key characteristics that had substantial pre-match imbalance (high pre-match standardized differences); thus this model included employer size, U.S. region, and calendar year of the index date. Evidence suggests that matching on baseline trends of outcome measures in interrupted time series studies closely approximates the effect estimates of randomized controlled trials.[10] We therefore included the key baseline outcomes of high- and low-severity emergency department visit and cost trends measured on a triannual basis (i.e., every 4 months) as count of visits or sum of standardized costs per 4-month period. Our final group included 23,493 HDHP members with diabetes and 192,842 matched controls.

III. Covariates

To estimate comorbidity, we applied the Adjusted Clinical Group (ACG) algorithm to members' baseline period. The algorithm uses age, gender, and ICD-9-CM codes to calculate a morbidity score and the average of the reference population is 1.0.[11] Researchers have validated the index against premature mortality.[12]

To derive proxy demographic measures, the data vendor linked members' most recent residential street addresses to their 2000 US Census block group.[13] Census-based measures of socioeconomic status have been validated[14,15] and used in multiple studies to examine the impact of policy changes on disadvantaged populations.[16-18] Income categories were based on living in neighborhoods with below-poverty-levels of <5%, 5%-9.9%, 10%-19.9%, and >=20%. We used a similar approach to categorize education levels (neighborhood residence with below-high-school education levels of <15%, 15%-24.9%, 25%-39.9%, and >=40%).[13-18] We defined low- and high-income as residence in neighborhoods with below-poverty levels of ≥10% and <10%, respectively. We also created lowest- and highest-income subgroups with corresponding below-poverty level cutoffs of ≥20% and <5%.

Although the 2000 Census variables might seem less than ideal given that our data ended in 2012, this remains the best source of area level income data for this type of large database analysis. The 2010 Census did not capture income and poverty levels because the US Census transitioned those questions to the American Community Survey (ACS). The ACS income and poverty variables are unfortunately problematic, as acknowledged by the Census Bureau itself[2,3] and confirmed by multiple reports and academic studies.[2-7] A recent examination concluded: "*The margins of error on ACS census tract-level*

data are on average 75 percent larger than those of the corresponding 2000 long-form estimate. The practical implications of this increase is that data are sometimes so imprecise that they are difficult to use."[7] In addition, The Committee on National Statistics' noted, "*Although research will be needed to evaluate income measurements across surveys, it is likely that the ACS will prove to be a relatively crude instrument for measuring income and poverty… Also, the "rolling" nature of the ACS may create measurement problems*."[5]

We classified members as from predominantly white, black, or Hispanic neighborhoods if they lived in a census block group (geocoding) with at least 75% of members of the respective race/ethnicity. We then applied a superseding ethnicity assignment if members had an Asian or Hispanic surname,[19] and classified remaining members as from mixed race/ethnicity neighborhoods. This validated approach of combining surname analysis and census data has positive and negative predictive values of approximately 80 and 90 percent, respectively.[20]

SUPPLEMENTARY DATA

IV. Analyses by Morbidity Level

Supplementary Table 3. Emergency department visits, hospitalizations, and total health care expenditures, overall and among other HDHP subgroups of interest, one year before and after a HDHP switch compared with contemporaneous control group members.

	Annual Rates[1]				Change in HDHP vs Control Group, Follow-up vs					
	HDHP Group		Control Group		Absolute			Relative, %		
	Baseline	Follow-	Baseline	Follow-	Estimate (95% CI)			Estimate (95% CI)		
High-morbidity (n=7299 HDHP and [illegible][2]										
Emergency Department Visits, per 1000	537.2	466.0	525.1	454.0	ND			ND		
Low-severity,[3] per 1000 Members	134.3	141.1	134.3	140.0	1.2	(-4.0,	6.3)	0.8%	(-2.9%,	4.5%)
High-severity,[3] per 1000 Members	61.3	59.9	61.3	57.4	2.5	(1.1,	3.9)	4.4%	(1.8%,	6.9%)
Hospitalizations, per 1000 Members	378.5	266.0	347.6	254.5	-19.3	(-42.6,	3.9)	-6.8%	(-14.5%,	1.0%)
Direct Admissions, per 1000 Members	223.7	142.4	202.7	140.0	**-18.6**	**(-32.2,**	**-5.0)**	**-**	**(-19.3%,**	**-3.8%)**
Total Expenditures, $ per Member	15754.6	13794.9	15387.6	13983.1	**-**	**(-651.7,**	**-**	**-3.9%**	**(-4.5%,**	**-3.2%)**
Lower-morbidity (n=15,292 and 121,919										
Emergency Department Visits, per 1000	114.7	200.4	114.7	210.8	**-10.4**	**(-14.4,**	**-6.5)**	**-5.0%**	**(-6.8%,**	**-3.1%)**
Low-severity,[3] per 1000 Members	37.7	63.2	37.4	68.2	**-5.0**	**(-6.1,**	**-3.8)**	**-7.3%**	**(-8.9%,**	**-5.6%)**
High-severity,[3] per 1000 Members	7.7	27.2	7.7	26.7	0.5	(-0.1,	1.2)	2.0%	(-0.5%,	4.6%)
Hospitalizations, per 1000 Members	20.6	81.9	24.0	85.3	ND			ND		
Direct Admissions, per 1000 Members	14.4	42.9	16.5	46.2	**-2.0**	**(-3.9,**	**-0.2)**	**-4.5%**	**(-8.5%,**	**-0.6%)**
Total Expenditures, $ per Member	5087.0	6349.3	5256.5	6712.8	**-10.4**	**(-14.4,**	**-6.5)**	**-5.0%**	**(-6.8%,**	**-3.1%)**

Abbreviations: HDHP, high-deductible health plan; ND, not detected. [1]All rates and changes account for differing baseline trends between HDHP and control group members and are estimated with marginal effects methods using parameters from aggregate-level segmented regression analysis of cumulative interrupted-time-series data that were adjusted for age, gender, race/ethnicity, education level, poverty level, US region, ACG score, employer size, and calendar month of the index date. [2]Adjusted Clinical Groups (see text) score of ≥3. [3]See manuscript for definition of low- and high-severity emergency department visits. [4]Adjusted Clinical Groups score of <2.

SUPPLEMENTARY DATA

Supplementary Table 4. Proxy adverse health outcomes of high-severity emergency department visit expenditures and high-severity hospitalization days (and low-severity outcomes for comparison), overall and among other HDHP subgroups of interest, one year before and after a HDHP switch compared with contemporaneous control group members.

	Annual Rates per 1000 Members[1]				Change in HDHP vs Control Group, Follow-up vs Baseline[1]					
	HDHP Group		Control Group		Absolute			Relative, %		
	Baseline	Follow-up	Baseline	Follow-up	Estimate (95% CI)			Estimate (95% CI)		
High-morbidity (n=7299 HDHP and 63,274 Control)[2]										
High-severity ED Visit Expenditures,[3] $ per Member	476.8	400.5	467.0	362.2	**38.2**	**(20.6,**	**55.8)**	**10.6%**	**(5.4%,**	**15.7%)**
High-severity Hospitalization Days,[3] per 1000 Members	180.2	163.7	156.0	145.4	-5.9	(-19.5,	7.7)	-3.5%	(-11.4%,	4.4%)
Low-severity ED Visit Expenditures,[3] $ per Member	413.9	380.8	384.2	363.8	**-33.9**	**(-66.8,**	**-1.0)**	**-8.2%**	**(-15.7%,**	**-0.6%)**
Low-severity Hospitalization Days,[3] per 1000 Members	123.3	115.4	114.2	98.2	8.0	(-4.8,	20.8)	7.5%	(-5.1%,	20.0%)
Lower-morbidity (n=15,292 and 121,919 Control)[4]										
High-severity ED Visit Expenditures,[3] $ per Member	23.9	160.4	21.2	155.4	2.3	(-3.2,	7.8)	1.4%	(-2.1%,	5.0%)
High-severity Hospitalization Days,[3] per 1000 Members	3.9	52.8	2.7	48.8	2.8	(-0.8,	6.4)	5.7%	(-1.8%,	13.2%)
Low-severity ED Visit Expenditures,[3] $ per Member	45.0	134.3	45.9	133.9	0.3	(-2.8,	3.5)	0.3%	(-2.1%,	2.6%)
Low-severity Hospitalization Days,[3] per 1000 Members	0.0	24.4	0.3	26.4	**-2.0**	**(-2.7,**	**-1.4)**	**-7.6%**	**(-10.0%,**	**-5.2%)**

Abbreviations: HDHP, high-deductible health plan; ED, emergency department. [1]All rates and changes account for differing baseline trends between HDHP and control group members and are estimated with marginal effects methods using parameters from aggregate-level segmented regression analysis of cumulative interrupted-time-series data that were adjusted for age, gender, race/ethnicity, education level, poverty level, US region, ACG score, employer size, and calendar month of the index date. [2]Adjusted Clinical Groups (see text) score of ≥3. [3]See manuscript for definition of low- and high-severity emergency department visits. [4]Adjusted Clinical Groups score of <2.

SUPPLEMENTARY DATA

References

1. Iacus SM, King G, Porro, G. Multivariate Matching Methods That are Monotonic Imbalance Bounding. *Journal of the American Statistical Association.* 2011;493(106):345-361.
2. Iacus SM, King G., Porro G. Causal Inference Without Balance Checking: Coarsened Exact Matching. *Political Analysis.* 2011.
3. S. I, G K, G P. CEM: Coarsened Exact Matching Software. https://gking.harvard.edu/cem. Accessed 19 May, 2017.
4. Schreyogg J, Stargardt T, Tiemann O. Costs and quality of hospitals in different health care systems: a multi-level approach with propensity score matching. *Health Econ.* 2011;20(1):85-100.
5. Wharam JF, Zhang F, Landon BE, LeCates R, Soumerai S, Ross-Degnan D. Colorectal Cancer Screening in a Nationwide High-deductible Health Plan Before and After the Affordable Care Act. *Medical care.* 2016;54(5):466-473.
6. Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *Journal of clinical epidemiology.* 1989;42(4):317-324.
7. D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med.* 1998;17(19):2265-2281.
8. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med.* 2007;26(1):20-36.
9. Coca-Perraillon M. Local and Global Optimal Propensity Score Matching. 2007; http://www2.sas.com/proceedings/forum2007/185-2007.pdf. Accessed 19 December, 2009.
10. St.Clair T, Cook TD, Hallberg K. Examining the Internal Validity and Statistical Precision of the Comparative Interrupted Time Series Design by Comparison With a Randomized Experiment. *American Journal of Evaluation.* 2014;35(3):311-327.
11. *The Johns Hopkins ACG Case-Mix System Reference Manual, Version 7.0.* Baltimore, MD: The Johns Hopkins University; 2005.
12. Reid RJ, Roos NP, MacWilliam L, Frohlich N, Black C. Assessing population health care need using a claims-based ACG morbidity measure: a validation analysis in the Province of Manitoba. *Health Serv Res.* 2002;37(5):1345-1364.
13. U.S. Bureau of the Census. *Geographical Areas Reference Manual*, Washington, D.C., U.S. Bureau of the Census. 1994.
14. Krieger N. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. *American Journal of Public Health.* 1992;82(5):703-710.
15. Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures--the public health disparities geocoding project. *American Journal of Public Health.* 2003;93(10):1655-1671.

16. Trivedi AN, Zaslavsky AM, Schneider EC, Ayanian JZ. Relationship between quality of care and racial disparities in Medicare health plans. *Jama.* 2006;296(16):1998-2004.
17. Trivedi AN, Rakowski W, Ayanian JZ. Effect of cost sharing on screening mammography in medicare health plans. *The New England journal of medicine.* 2008;358(4):375-383.
18. Selby JV, Fireman BH, Swain BE. Effect of a copayment on use of the emergency department in a health maintenance organization. *New England Journal of Medicine.* 1996;334(10):635-641.
19. Ethnic Technologies. http://www.ethnictechnologies.com/index.html.
20. Fiscella K, Fremont AM. Use of geocoding and surname analysis to estimate race and ethnicity. *Health services research.* 2006;41(4 Pt 1):1482-1500.