APPENDIX A: Description of Data and Data Cleaning

Appendix A1: Datasets and Sources

Health Care Cost Institute (HCCI) Data: Our paper draws on data from the Health Care Cost Institute (HCCI). The HCCI data include claims from beneficiaries with employer-sponsored coverage from Aetna, Humana, and UnitedHealthcare. More details on HCCI can be found at www.healthcostinstitute.org.

The data include claims for individuals with fully-insured and self-insured plans that receive employer-sponsored insurance.[1] This includes insurance products in the national, large, and small group markets. The data cover 27.6 percent of individuals in the US with employer-sponsored insurance. The data begin with sheets of membership data, inpatient facilities data, outpatient data, physician data, and pharmacy data. We use these to construct our inpatient and procedure samples. A sample hip replacement case constructed from these claims is posted online at http://healthcarepricingproject.org/sites/default/files/papers/sample_hip_claims.xlsx. This illustrates how we aggregate claims up to the case level and calculate a price.

While the HCCI data include more than forty million covered lives per year (see Appendix Table I in the body of the paper), the data are from health insurance claims for individuals with health care coverage from Aetna, Humana, or UnitedHealthcare. While these are three of the largest five health insurers in the US, we do not have claims from Blue Cross Blue Shield (BCBS) health insurers. BCBS is an association of 38 for-profit and not-for-profit health insurers in the US who purchase a license to use the BCBS name. We use membership data from our database and compare it to coverage rates in the American Community Survey and the Census Bureau's Small Area Health Insurance Estimates (SAHIE) to estimate the coverage of our three insurers at the state and county level. We also use data from the HealthLeaders Interstudy database to estimate the share of lives BCBS insurers cover by county. We use this information to show that our results are robust to areas with different levels of HCCI and BCBS coverage (See Appendix F).

The most prominent alternative source of private health insurance claims data is the MarketScan database from Truven Health Analytics. MarketScan data include claims for individuals with health insurance from a number of large employers and also some smaller employers (although it seems that the MarketScan coverage for smaller employers is substantially lower than their coverage for larger employers). Most previous research using the MarketScan data to analyze health spending has relied on only the claims for individuals employed by large firms. We use the HCCI data to analyze claims for individuals employed in small, medium, and large firms. Using the HCCI data allows us to look at a substantially larger population than has been analyzed using the MarketScan data. Chernew et al. (2010) report that the MarketScan data contain between 16.9 million and 22.9 million covered lives per year between 1996 through 2006. By contrast, the HCCI data contain between 42 and 46 million lives per year (see Appendix Table I).

[1] With fully-insured plans, the insurer pools and bears risk. With self-insured plans, the firm pays all insurance claims themselves and relies on insurance companies for administrative services.

While the MarketScan database is useful for many research applications, it has drawbacks for the type of analysis we undertake in this project. First, the MarketScan database does not contain hospital IDs and sub-three digit geographic identifiers. A unique hospital identifier is necessary so that we can merge in hospital characteristics and, more importantly, analyze price variation within and between providers. With HCCI, we can merge on hospital characteristics, identify individual hospitals, and merge in local characteristics at the zip code level. Second, MarketScan has very thin coverage in a number of markets. For example, while the smallest HRR in the HCCI data has 2,932 unique individuals, MarketScan includes HRRs with fewer than two hundred individuals.

In addition to the core HCCI data, we merge on a number of other datasets listed below.

American Hospital Association Annual Survey: We obtain data on hospital characteristics from the American Hospital Association (AHA) annual survey. More information on the AHA survey data can be obtained from: http://www.ahadataviewer.com/book-cd-products/AHA-Survey/. The survey polls hospitals on characteristics, staffing, technology, finances, and other information and has been running since 1946. We use the AHA data to create our technology measures and measures of hospital market structure.

American Community Survey Data: We use data on the percentage of working age (18-64) adults with employer-based health insurance coverage by county from the American Community Survey conducted by the US Census Bureau, https://usa.ipums.org/usa/acs_healthins.shtml.

American Hospital Directory Data: We use data on hospitals Medicare activity that we obtained from the American Hospital Directory (AHD). The AHD is a for-profit data vendor that sells cleaned Medicare claims data derived from the Medicare Provider Analysis and Review limited access database. This includes claims records for 100% of Medicare fee-for-service inpatient claims. Details on the AHD data can be found at www.ahd.com.

Census Data: Data on the number of uninsured lives by county, lives privately insured per county, and median household income come from the US census. See: http://www.census.gov/did/www/sahie/ and http://www.census.gov/did/www/saipe/index.html.

Dartmouth Data: We use data on Medicare spending per HRR that we downloaded from the Dartmouth Atlas. Full details on the Dartmouth Atlas Medicare data can be obtained from: http://www.dartmouthatlas.org.

FactSet Research Systems: These reports provide a roster of merger and acquisition (M&A) activity across industries and include the names of firms involved in transactions and the date of transactions. We used the database to find hospital mergers. The data are accessible with a subscription at: https://www.factset.com/data/company_data/mergers_acq

HealthLeaders Interstudy Data: The HealthLeaders Interstudy database, available for purchase from the Decision Resources Group, includes the count of individuals enrolled, by county, by insurer in the small, medium, and large group markets. The data include coverage of the self-insured and fully-insured market. See: decisionresourcesgroup.com.

Irving Levin Associates' Health Care Services Acquisition Reports: These reports provide a roster of M&A activity in hospitals, managed care companies, physician medical groups, rehabilitation centers, labs, and behavioral health groups. We used reports for 2007 to 2011 to identify the hospital mergers that we include in this analysis. The reports can be purchased from: https://products.levinassociates.com/downloads/har-2017/

Medicare Quality Scores: We use data on hospital quality obtained from data.medicare.gov. The data include quality scores drawn from both Medicare and private claims data. The data can be downloaded from: https://data.medicare.gov/data/hospital-compare. The quality scores used were developed by the Agency for Health Care Research and Quality (AHRQ).

Securities Data Company (SDC) Platinum: This database provides a historical transaction database including a roster of hospital mergers. The data are accessible with a subscription via: https://financial.thomsonreuters.com/en/products/data-analytics/market-data/sdc-platinum-financial-securities.html.

U.S. News & World Report Rankings: We obtained rankings of hospitals printed in the US News and World Report from 2007 –2011. Some data were obtained from online rankings. For some years, we obtained the physical copy of the printed magazine issues.

Appendix A2: Identifying Hospitals Using National Plan and Provider Enumeration System Identifiers

Single hospitals can be assigned multiple National Plan and Provider Enumeration System Identifiers (NPI) because different wings of the hospitals and different units can each have their own NPI (e.g. a hospital's radiology service could have a separate NPI to its Emergency Room). To address this issue, we made a crosswalk that consolidates providers' multiple NPIs into a single, master NPI. We use the master NPI to merge on data from the AHA and Medicare. To consolidate NPIs, we undertake the following steps:

1. Compile all variations of AHA ID/hospital name/address/city/state/ZIP Code in the 2000-2011 AHA survey data, retaining the row for the latest year.
2. Add NPI from the AHA survey files, beginning with the most recent year.
3. Make sure there is only one NPI per AHA ID. If more than one AHA ID have the same NPI, look up in the CMS NPI Registry to resolve the discrepancy.
4. Check all NPIs in the CMS NPI Registry to make sure they are valid and accurate. Remove invalid NPIs.
5. Look up hospitals in the NPI Registry that do not have an NPI in AHA by name and address. Attach NPI to the AHA file when a match is found.
6. Extract all organizational rows from the CMS NPI Registry where primary taxonomy code is for a hospital (287300000X, 281P00000X, 281PC2000X, 282N00000X, 282NC2000X, 282NC0060X, 282NR1301X, 282NW0100X, 282E00000X, 286500000X, 2865C1500X, 2865M2000X, 2865X1600X, 283Q00000X, 283X00000X, 283X00000X, 283XC2000X, 282J00000X, 284300000X) or hospital unit (273100000X, 275N00000X, 273R00000X, 273Y00000X, 276400000X).
7. Match AHA compiled address file to the hospital NPI file on NPI. Add AHA number to the hospital NPI file and mark the NPI as "PRIMARY" NPI for that hospital.
8. Match remaining rows in the hospital NPI file according to the following hierarchy:
    1. Organization name, address1, city, state, ZIP Code
    2. Address1, city, state, ZIP Code, similar organization name

    3. Other organization name, address1, city, state, ZIP Code
    4. Address1, city, state, ZIP Code, similar other organization name
    5. Address, city, state, ZIP Code, different name (validated name changes via web search)[2]
    6. Organization name, similar address1, city, state, ZIP Code[3]
    7. Other organization name, similar address1, city, state, ZIP Code
    8. Similar organization name, similar address1, city, state, ZIP Code
    9. Similar other organization name, similar address1, state, ZIP Code
    10. Medicare number, city, state, ZIP Code
9. When a match is found, append AHA ID and "PRIMARY" NPI.
10. Some hospitals in the NPI Registry were not in the AHA survey data files. For these hospitals, we pick one NPI as "PRIMARY" and, using the match steps outlined above, add an "X" to the AHA ID column and append the "PRIMARY" NPI to all matched rows.
11. We also consolidated NPIs to ZIP codes. To do so, we:
    1. Sort file by ZIP Code, primary taxonomy code, address1
    2. Where more than one "PRIMARY" NPI exists within a ZIP Code for the same organization name and primary taxonomy, change all rows to the "PRIMARY" NPI associated with the AHA ID.
    3. Where more than one "PRIMARY" NPI exists within a ZIP Code for the same organization name and primary taxonomy but none of the rows is associated with an AHA ID, double check against the AHA file. If no match is found, consolidate the rows to one single "PRIMARY" NPI.

<u>Appendix A3: Constructing a consistent hospital-level panel from the AHA Data</u>

When hospitals merge, the AHA Survey will often consolidate two hospital IDs into a new single ID. While this does not affect our measure of hospital prices (since those are generated from the HCCI data), it does delete observations from the AHA data. This creates two issues. First, according to the AHA data, the count of AHA hospital sites (as opposed to systems) decreases over time. This is caused mechanically by mergers, which reduce the numbers of IDs. Second, because we measure prices for hospital sites, AHA characteristics that we use as control variables are only available at the more aggregated level of the consolidated sites. While most of our control variables are categorical (e.g. whether a hospital is a teaching facility), some are continuous measures (e.g. hospital beds, the count of Medicare discharges per year, and the count of Medicaid discharges per year).

A good example of this issue is that after their merger, the IDs for New Britain General Hospital in New Britain, CT and Bradley Memorial Hospital in Southington, CT are consolidated into a new ID number for the Hospital of Central Connecticut in 2006. In the AHA Survey data the IDs for "New Britain General" and "Bradley Memorial" vanish from the survey in 2006 and a new hospital ID for "Hospital of Central CT" appears in the same year.

[2] Because there can be hospitals within hospitals (e.g., specialty or children's hospital on one floor of a general hospital), all of these occurrences were manually validated to ensure that the correct hospital was identified.
[3] Suburb names are occasionally used in addresses (e.g., Brentwood vs. Los Angeles). If the address1, state, and ZIP Code matched but the city name differed, this was still considered a valid match at each level.

This is a standard problem in firm-level analysis. A firm is composed of a number of establishments and often data are only available at the higher firm-level (e.g. Compustat). When two firms merge information is often only available at the aggregate consolidated level and not for the individual firms (even when they are still run as separate businesses). A standard approach to this problem is to freeze the organizational structure at a point of time, so the researcher can analyze a consistent set of firm sub-units (or at least until they exit). We perform an analogous exercise for hospital sites.

In order to maintain the information at the more disaggregated level we "undo" the site-level consolidation in AHA after 2001 by (i) maintaining the original (vanished) ID at the site level in the year the consolidation occurs and for all years afterwards; (ii) remove the new consolidated ID from the data in all years after it occurs.[4] We then construct a new master hospital system ID. The challenge that arises from "undoing" this consolidation of IDs is we do not know the correct bed count (and other observables) at the hospital site-level after consolidation.

We address this by imputing the information at the consolidated level to the site level for all continuous variables for these hospitals in the following manner. Consider the following example of imputing hospital beds. Let two separate hospitals have distinct IDs A and B at time $T\text{-}1$. Assume that hospitals A and B merge at time T and become hospital C (hospital C may have already been in existence at T-1 or may be a new hospital created from the merger of A and B at time T). The merged hospital is given the ID C and the IDs for A and B cease to exist. Let b_t^h be the number of beds at hospital h at time t where $h \in \{A,B,C\}$ and $t \in \{2001, 2002, \ldots, 2014\}$. Let $w^h = \frac{b_{T-1}^h}{\sum_{h \in \{A,B\}} b_{T-1}^h}$. w^h is hospital h's share of the total number of beds between hospitals A and B at time $T\text{-}1$. If $\frac{\left|b_T^C - \sum_{h \in \{A,B\}} b_{T-1}^h\right|}{\frac{b_T^C + \sum_{h \in \{A,B\}} b_{T-1}^h}{2}} \leq 0.2$, then we assume hospital h's bed total is $w^h b_t^C$ for all t in which hospital C exists in the AHA Survey. Otherwise, we assume hospital h's bed total is b_{T-1}^h for all t in which hospital C exists.

In other words, if the percentage difference between the total number of beds at A and B in $T\text{-}1$ and the number of beds of the consolidated hospital ID in time T is less than or equal to 20 percent, then we impute hospital A's bed count to be its share of the total beds at A and B at time $T\text{-}1$ (w^A), multiplied by the consolidated hospital's total number of beds (b_t^C) for all years that hospital C exists in the AHA Survey. If this percentage difference is greater than 20 percent, then we assign hospital A the bed total it has at time $T\text{-}1$ to all the years in which hospital C exists (from time T forward).[5]

We carry out this same imputation procedure for the share of Medicare and Medicaid discharges using the above methodology.

[4] In some cases, the merger is recorded using the aggregation of an acquired hospital into an existing AHA ID. In these cases, the procedure is the same except we do not delete observations for the acquiring hospital.

[5] We choose a threshold because if the difference is large then it indicates that the merged hospital is undergoing a large restructuring, so this casts doubt on the assumption that the relative size of original entities is stable. 20 percent is an arbitrary threshold, of course, but the results are robust to other reasonable thresholds.

__Appendix A4: Defining the Inpatient and Procedure Pricing Samples__

The inpatient sample in our data includes all inpatient claims aggregated to the level of a single hospital admission (which we call a case), each of which has a unique DRG. The procedures we use are defined using combinations of ICD9 codes and DRGs. In the case of MRIs, we identify cases using CPT-4 codes. The specific codes we use to define samples include:

Coding Definitions for the Seven Procedure Samples

Procedure	**ICD9**	and	**MS-DRG**	or	**CPT-4**
Hip Replacement	8151		470		
Knee Replacement	8154		470		
Cesarean Section	741		766		
Vaginal Delivery	7359		775		
PTCA	0066		247		
Colonoscopy	V7651 (CM)				
MRI					73721

For hip and knee replacements, we limit our analysis to individuals between forty-five and sixty-four years of age. For vaginal deliveries and cesarean sections, we limit our analysis to delivering mothers who are between the ages of twenty-five and thirty-four. In order to be included, an MRI case must be a single-line facility claim and we must observe a separate physician payment for the reading of the MRI. We do this to ensure that we are isolating the professional component (reading of the MRI) from the technical component (administering the scan). We also limit MRIs to those carried out on individuals who had no other hospital claims on the day that the MRI was provided and for whom the hospitalization was exclusively for the MRI. Similarly, for colonoscopies, we limit our analysis to individuals aged forty-five through sixty-four and only include hospital-based cases where nothing else was done to the patient that day and for which the colonoscopy was the reason for the trip to the hospital. We exclude colonoscopies where a biopsy was taken.

In order to minimize the impact of unusually complicated cases or clerical billing errors, we exclude cases above the 99th percentile of length-of-stay as well as cases where the price is below the 1st percentile or above the 99th percentile. In the inpatient sample, these restrictions are implemented by DRG.

Appendix Table III shows the impact on the number of hospitals and cases of the main selection criteria we use to derive our inpatient sample. After conditioning our data to cases delivered at hospitals that are registered with the AHA, we have 5,865,727 inpatient cases delivered at 4,326 facilities between 2008 and 2011. Excluding critical access hospitals drops our number of providers by 1,124 (26 percent), but only lowers the number of cases we observe by 51,349 (less than one percent). We further exclude three hospitals where we do not have data on Medicare activity. We then exclude all cases from 2007. This lowers our cases by 769,104 (13 percent) and number of hospitals by 10 (less than one percent). In order to have sufficient data at each hospital to calculate an inpatient price index, we exclude providers that had fewer than 50 cases per year. This drops 74,705 cases (1.5 percent) and 831 hospitals (26 percent).

Appendix A5: Construction of Price Fixed Spending and Quantity Fixed Spending Used in Section III.B.

We calculate Medicare and private spending per beneficiary where we fix quantities nationally (and only allow price variation to drive variation in spending) and fix prices (and only allow quantity variation to drive spending variation).

To do so, we first calculate inpatient spending per beneficiary for the privately insured and for Medicare recipients. Inpatient spending per beneficiary in HRR r (y_r) is a function of the quantity (q_r) of care provided and the price of care (p_r):

$$y_r = \frac{\sum_{h,d}(p_{h,d} q_{h,d})}{B_r},$$

where the price of DRG d at hospital h in HRR r is represented by $p_{h,d}$ and quantity is $q_{h,d}$ (we suppress the subscript r for economy of notation), B_r is the number of beneficiaries in HRR $r,$ and $\sum_{h,d}$ indicates summing across all DRGs in a hospital and the all hospitals in an HRR.

We compute counterfactuals to calculate the relative contributions of price and quantity to variation in inpatient spending. The first counterfactual we create is to fix all prices per DRG to be the same as the national average ($\bar{p}_d$) and then analyze spending variation. This allows us to identify the relative contribution that differences in the quantity of care provided across regions make to variation in spending per beneficiary. Spending per beneficiary calculated with national average prices is (where ~ indicates a counterfactual calculation):

$$\tilde{y}_r^{\bar{p}} = \frac{\sum_{h,d}(\bar{p}_d q_{h,d})}{B_r}.$$

The second counterfactual is to fix the quantity and mix of inpatient care delivered in each HRR to be the same as the national average mix and quantity of care ($\bar{q}_d$) and then analyze spending variation.[6] To do so, we calculate:

$$\tilde{y}_r^{\bar{q}} = \frac{\sum_{h,d}(\bar{q}_d p_{h,d})}{B_r}.$$

This allows us to identify the relative contribution that differences in price make to variation in spending per beneficiary across HRRs. These are, of course, purely accounting decompositions to gauge rough magnitudes, as quantity and price are both endogenously determined in the private sector.

Appendix Tables VII and VIII contain the results of these counterfactual calculations for individuals age 55 to 64 (Appendix Table VII) and individuals age 18 to 64 (Appendix Table VIII).

Appendix A6: Construction of Control Variables for Sections VI and VII

[6] To do so, we identify the mix of DRGs at a national level and set every HRR to have that mix of DRGs.

In our estimates of the relationship between market structure, mergers, and hospital prices in Sections VI and VII, we also include a range of additional hospital and local area controls. Below are descriptions of these additional measures.

Hospital Characteristics and Hospitals' Local Area Characteristics: In our cross-sectional and merger analysis, we include controls for hospital characteristics drawn from the AHA annual survey. These include: the number of hospital beds, ownership type (not-for-profit, for-profit, government), teaching status, and indicators for the technologies available at a hospital in a specific year. In addition, we link hospitals' zip codes to local area characteristics from the Census Bureau's Small Area Health Insurance Estimates and Small Area Income and Poverty Estimates, including the proportions of the population who are uninsured and the median income in the county where the hospital is located.

Technology Index: We follow Acemoglu and Finkelstein (2008) in using a count of hospital technologies offered by a hospital as recorded in the AHA survey data. The AHA data include binary indicators for whether a hospital has various technologies and services, such as computer-tomography (CT) scanners, electron beam computed tomography, or proton beam therapy. We sum the number of these technologies available at each hospital in each year.

Hospital Quality: To capture reputational quality, we include a yearly indicator for whether or not a hospital was ranked by the U.S. News & World Report as a top hospital. We indicate a hospital was ranked in the U.S. News and World Report if it was ranked as an overall top hospital or received a ranking as a top hospital for cancer care; gastrointestinal care; ear nose and throat; geriatric care; gynecology; cardiology; orthopedics; rheumatology; or urology. In total, from 2008 through 2011, the U.S. News & World Report ranked 192 hospitals in our sample in their annual "Best Hospital" rankings across clinical specialties and the overall ranking.

To measure clinical performance, we merge in data on hospital quality from https://data.medicare.gov/, which includes the hospital quality scores reported publicly on the CMS Hospital Compare webpage (https://medicare.gov/hospitalcompare). These include measures of patient safety, patient outcomes, and process measures of care captured from public and private claims data. We included quality scores for 2008 through 2011 for four measures: the percentage of heart attack patients given aspirin upon arrival to the hospital; the percentage of surgery patients given an antibiotic prior to surgery; the percentage of patients treated within twenty-four hours of surgery to prevent blood clots; and the 30-day risk adjusted mortality from heart attacks.[7] These are widely acknowledged measures of the quality of care and they are all available for hospitals in our sample from 2008 through 2011 (Yale Center for Outcomes Research and Evaluation 2013). We focus on these four clinical quality measures in the robustness analysis, but we also examine the effect of conditioning on all 41 quality measures. Note that we do not have CMS quality measures for 168 hospitals (7.5 percent) from our inpatient sample. As a result, we present analysis of these measures separately from our main analysis.

Medicare and Medicaid Activity: We include the Medicare base payment rate for hospitals, since this may proxy for hospital costs. This comes from annual Medicare Impact Files. We also

[7] For the technical descriptions of the measures of performance we used in this analysis, see http://www.medicare.gov/hospitalcompare/Data/Measures.html.

include data from the AHA on the share of hospitals' inpatient cases paid by Medicare and Medicaid each year.