The Health Systems and Provider Database
Technical Documentation
January 2021

METHODOLOGY FOR IDENTIFYING AND CLASSIFYING HEALTH SYSTEMS

I. Identification of Health Systems

We defined a health system to be a set of provider organizations that are jointly owned or managed. The set of commonly owned or managed providers must have met the following additional criteria:

- contain a minimum of one general acute care hospital, ten primary care physicians whose primary billing TIN is owned or managed by the system, and fifty total physicians primarily billing under a system TIN, and
- the minimum set of providers must be located within a single hospital referral region

The methodology we developed to empirically identify health systems is outlined below and then described in detail. Health systems were identified in each year using an internally developed algorithm implemented as SAS, R, and Stata code operating on a large number of input data sets. The methodology had five major steps:

Step 1: Create files of providers (e.g. hospitals, physicians, physician practices) actively delivering care in the United States.

Step 2: Identify the tax identification numbers of corporate organizations (e.g. chain home offices, foundations, holding companies, and corporate subsidiaries) that own or manage provider organizations. Group commonly owned corporate subsidiaries.

Step 3: Identify the tax identification numbers for hospitals and post-acute care facilities (from step 1), identify the owners/managers of these facilities, and combine these providers with their owning corporate organizations (from step 2).

Step 4: Identify owners/managers of physician practice organizations and add practice organizations to networks containing their owning or managing entities (e.g. corporate organizations, hospitals, etc.)

Step 5: Qualify networks as health systems if they meet the definitional requirements.

In the following paragraphs we describe each of these steps in detail and the input data sources.

Step 1: Create provider files.

Our **physician file** was created by combining data from the following sources: CMS Provider Enrollment and Chain Ownership System (PECOS) file, CMS Physician Compare, IQVIA physician file, CMS Medicare Data on Provider Practice and Specialty (MD-PPAS), traditional FFS Medicare claims, commercial claims data, CMS MAX Provider Characteristics (MAXPC) file, and extracts from state All Payer Claims Data (APCD). We defined a physician as a medical doctor (MD) or a doctor of osteopathy (DO) and identified physicians by their National Provider Identification (NPI). We relied on a set of primary data sources (i.e. PECOS, IQVIA, Medicare claims, commercial claims, MAXPC, APCDs) to restrict our dataset to clinicians

who are delivering care to patients at some point during the year. We used the NPPES to identify physicians separately from other clinicians whose NPIs appeared in our primary data sources (e.g. nurse practitioners). Data from the National Plan and Provider Enumeration System (NPPES), MD-PPAS, CMS Physician Compare, IQVIA, and commercial claims were combined to classify specialty for each physician. To code physician specialty uniformly, we developed a physician specialty taxonomy based on board certifications offered by the American Board of Medical Specialties (ABMS) and mapped specialty classifications in each of the input data sets to the taxonomy. We then combined all available specialty data for each physician and assigned each physician to one of 123 ABMS terminal specialties (details are available upon request). Physicians were classified as primary care if their primary specialty is family practice, general practice, pediatrics, geriatrics, or internal medicine with no medical or surgical specialty (excluding pediatric subspecialty).

We defined a physician practice as a legal entity that is fully or partially owned by physicians (e.g. sole proprietorship, partnership) or that employs physicians actively delivering care. To create the **physician practice file**, we combined data from PECOS, Medicare claims, IQVIA practice site file, and commercial claims.

We used two PECOS files to identify the practice TINs through which physicians submit claims. The PECOS reassignment file contains observations for every NPI-TIN combination for which the Medicare-certified physician (NPI) has reassigned his or her Medicare benefits (i.e. payments) to a provider organization (TIN). Physicians may reassign benefits to multiple provider organizations and these provider organizations may be physician practices (corporations or partnerships), hospitals, other health care facilities, and corporations (e.g. health systems, joint ventures). The PECOS enrollment associations file also contains observations for physician practices and provides information on their relationships with individuals (e.g. partner) and other organizations (e.g. owning or managing entities, billing agencies). Importantly it contains information on practice organizations established as sole proprietorships; these organizations are not included in the PECOS reassignment file. We used role codes in the PECOS enrollment associations file to identify these practices and to generate NPI-TIN observations. The NPI-TIN observations extracted from these two PECOS files were then checked against Medicare claims (carrier, outpatient, inpatient, PartD and MDPASS) data, commercial claims, CMS Physician Compare, and the list of physicians who have opted out of Medicare. Any NPI-TIN observations from PECOS that were billing claims (LCD, Carrier, OP, IP) in a previous year but not the current year were assumed to be inactive in the current year. NPT-TIN pairs which do not have any previous claims but are in PECOS were assumed to be valid. In addition, there are NPI-TIN observations present in Medicare claims data that do not have corresponding observations in any PECOS file; we believe they are missing from PECOS data because these relationships pre-dated the creation of certain PECOS files, were not required to formally enroll, and have been grand-fathered into the system. Any TIN observed in Medicare Carrier or LCD claims data that are not also present in the PECOS data were included as physician practices.

IQVIA data does not contain TINs for provider organizations. For each active physician (NPI) in our physician database that is not observed in one of our claims databases, we searched the set of practice site-NPI combinations in IQVIA data and included the matched practice sites as observations in our physician practice database. This screening process allowed us to include physician practices whose members are actively serving patients but who do not bill one of the payers from whom we have claims data. Most of the physicians observed in IQVIA and not in claims data are pediatricians.

We defined an acute care hospital to be a facility with at least 6 beds available for patients receiving inpatient care for acute medical conditions. We drew on three primary data sources to create our **short-**

term acute care hospital file: CMS provider of services (POS) file, IQVIA hospital file, and American Hospital Association (AHA) survey data. None of these data sources are comprehensive and each has limitations. In addition, there is no observation ID that is used consistently in all three sources. A couple of limitations are worth noting. The POS file is generated from administrative data and contains observations on hospitals that are closed or have changed their primary use and no longer provide acute care services (e.g. hospitals that have converted to post-acute care facilities). We used the hospital tracking files available through the Dartmouth Atlas website (https://atlasdata.dartmouth.edu/static/supp_research_data#hospital-research-data) combined with web-based research to address deficiencies in the CMS POS data. CMS allows multiple hospitals with the same owner to report in consolidated form, as a single entity, under one CMS Certification Number (CCN). All information reported in the POS file with the exception of location (e.g. count of beds, services available) refer to the combined set of hospitals reporting under a single unique CCN (usually the CCN of the larger hospital); the address listed for each CCN is the address of the main hospital. The AHA similarly allows a "parent" hospital to fill out one survey for the combination of the parent and "unit" hospitals. In contrast, IQVIA data contain unique observations for each hospital facility but sometimes list units within a hospital (e.g. a pediatric center) as a separate hospital facility. Finally, in each of these hospital datasets, the dominant type of services provided at the hospital (e.g. general acute care, pediatric care, orthopedic surgery, cancer care) is not coded in a comparable fashion.

To overcome the limitations of these individual hospital databases, and to ensure that our hospital data would be comprehensive (i.e. that it would include a unique observation for every short term acute care inpatient facility), we combined data from all three sources using a novel approach. First, because address was the only comparable field included in all three sources, we geocoded the address of each facility included in each hospital database. Second, we grouped observations into location IDs based on their proximity to one another, taking into account any available cross source references (e.g. AHA reports CCNs for some hospitals) and address string comparisons. Third, we created a new hospital service type variable based on the source's original coding of primary service type and string searches of the hospital name. Fourth, within location ID, we combined observations from different sources with the same hospital service type. Fifth, we manually reviewed cases of multiple hospital facilities with the same services type at a given location ID. This set of steps resulted in a hospital file containing a unique observation for each inpatient hospital facility located at a unique address and providing a unique primary service. We created a hospital ID that is consistently coded over time and linked that to the ID variables in each of the source datasets (e.g. CCN, AHA ID, IQVIA ID). The variable for primary service type may take one of the following values: general acute care (medical and surgical), children's, psychiatric (includes hospitals treating substance use disorders), heart, women's, cancer, orthopedic, geriatric, other specialty.

We defined post-acute care (PAC) facilities to include skilled nursing facilities, rehabilitation facilities, home health agencies, and hospice. Because long term acute care facilities are qualitatively different from short term acute care hospitals and may see some of the same types of patients as PAC facilities, we included these providers in the **PAC provider file**. Data on these facilities was extracted from the CMS POS file.

Step 2: Identify and Group TINs of Health Care Corporations and their Subsidiaries.

Health systems can be defined empirically as sets of TINs that are a mix of provider TINs and corporate (non-provider) TINs. Many large health care systems have a complex corporate organization in which the constituent provider organizations are owned by corporate subsidiaries which in turn are owned or

governed by a single corporate entity. Most of the TINs in these corporate hierarchies are non-provider corporate TINs.

Occasionally, provider organizations have multiple owners (e.g. they are organized as a joint venture corporation or partially or fully owned by private equity) or affiliations with multiple health care corporations. Multiple owners and affiliations present special challenges to identifying distinct health systems. To avoid combining two or more health systems connected by jointly owned or affiliated providers, we began building our health systems with "wholly-owned" TINs. Wholly-owned TINs have a single owner and can therefore be grouped together to form systems without inadvertently combining otherwise distinct networks of providers. Health system corporate TINs are typically wholly-owned, but provider TINs may be jointly owned or have affiliations with multiple health care corporations that are difficult to distinguish from ownership relationships in PECOS data and other TIN-based datasets.

We used a "top-down" empirical strategy to build networks of wholly owned TINs. The first step was to identify a set of wholly-owned TINs in PECOS data. Chain home offices (CHOs) in PECOS are defined as "[groups] of two or more providers under common ownership or control" (42 CFR 421.404). Based on this legal definition, we are fairly certain that these entities are wholly-owned. The PECOS Chain Home Office Addresses File contains the postal addresses for the PECOS chain home offices. We grouped the TINs corresponding to Chain Home Offices with the same address into a single network.

Academic health systems are another group of wholly-owned TINs. Each school is separately licensed or approved by the state and has its own board of trustees (U.S. Department of Education, "Organization of U.S. Education: Tertiary Institutions"). We identified their TINs in PECOS by searching for the words "College" and "University" in providers' legal business names. Since this text-based search yields providers that are not part of academic health systems (e.g., the multi-specialty group University Physicians Group located in New York state), we manually screened for false positives and dropped them from our list of TINs. We then grouped TINs by college, university, or university systems and added these sets of TINs to the list of chain home office TINs.

The second step in creating networks was to combine wholly-owned TINs with the same ownership or managerial control. Some of the wholly-owned TINs on our list, particularly the PECOS chain home offices, are subsystems that together form a single health system (e.g. Ascension Health and Alexian Brothers Health System). To group wholly-owned TINs we used multiple data sources to create a dataset of TIN-TIN pairs related by common ownership or managerial control. The input data sources for this TIN-TIN dataset are:

- *PECOS Chain Home Office Addresses File*. This file contains the postal addresses for the PECOS chain home offices. Some systems and subsystems list the same address(es), which we exploited to put related organizations in common networks.
- *IRS Business Master File*. This file lists related organizations that file a single common tax return. We used these data to group together wholly-owned TINs that are part of the same filings.
- *IRS 990 Filings for Tax-Exempt Entities sourced from both a proprietary data set prepared for us by Guidestar (now Candid) and filings hosted by Amazon Web Services (AWS)*:
    - Main Form with "Doing Business As" names and website addresses. We grouped together wholly-owned TINs with the same "Doing Business As" names and/or website addresses.
    - Schedule R with filers' related organizations and their direct controlling entities. We grouped together wholly-owned TINs that appear as filers and/or related organizations with hospitals and subsidiaries listed in Schedule R filings as directly controlled entities. Based on the definition of related organizations and direct controlling entities, we believe

TIN pairs constructed from this Schedule R data describe ownership and managerial control and do not constitute mere affiliations. We believe the hospitals that appear these TIN pairs are also wholly-owned entities (because they appear as directly controlled entities of a wholly-owned TIN).

- *Annual SEC10-K Filings.* Exhibit 21 lists each filer's subsidiaries. We used these data, compiled by the non-profit organization CorpWatch, to group together wholly-owned TINs.
- *S&P Capital IQ M&A Transactions.* This file contains many consummated M&A Transactions during our time period. We used these data to group together wholly-owned TINs and/or take them apart to correct for time lags in our other data sources.
- *Irving Levin Associates Deal Search Online Database and Health Care Services Acquisition Reports, 2010 - 2018, editions 17 - 25 (www.healthcaremanda.com).* These data contain announced M&A Transactions, some of which have been consummated. We identified consummated transactions between wholly-owned TINs on our list and used these data (like the S&P Capital IQ M&A Transactions above) to correct for lags in our other data sources.
- *Hand coding.* Through this process, we identified errors in the data that we fixed with hand coding.

From these sources we created a single dataset of wholly-owned TIN pairs that are related through common ownership or managerial control relationships. With these pairwise relations, we used R's igraph package to create mutually exclusive groups of TINs. Combining these wholly-owned TINs allowed us to avoid the incorrect combination of corporations linked through joint ventures and other partial ownership arrangements. It also helped us combine systems comprised of several subsystems. At the end of this step, we had a file that grouped together wholly owned TINs that are commonly owned or managed.

Step 3: Identify TINS for hospitals and PAC facilities and add these providers to networks of owning entities.

To add hospitals and post-acute care facilities to the wholly-owned networks created in step 3, we first identified the TINs used by these facilities for billing and reporting. In POS data, these facilities are identified using CMS Certification Numbers (CCNs). We used data from the PECOS Medicare ID file to find PECOS enrollment IDs for each facility CCN and then used the PECOS main provider file to link each enrollment ID to a facility TIN. Hospital and PAC CCNs may be associated with more than one TIN.

The second step was to create a dataset of TIN-TIN pairs that include at least one hospital TIN and where the TIN pairs are related by common ownership or managerial control (similar to the file of TIN-TIN pairs we created in step 3 to combine wholly-owned TINs into networks). The input data sources for this hospital TIN-TIN dataset are:

- IRS 990 Forms sourced from Guidestar (now Candid) and AWS.
    - Schedule A with filers' supported organizations. This file contains the supported organizations for IRS 990 filers that are 509(a)(3) organizations.
    - Schedule H with hospital facilities. This file contains the hospital facilities managed and controlled by IRS 990 filers who are hospital corporations.
    - Schedule R with (i) TIN pairs between the IRS 990 filer and its related organizations and (ii) TIN pairs between related organizations and their direct controlling entities. We limited TIN pairs to those between hospitals as well as those between our list of wholly-owned TINs (from Step 3) and hospitals.
- IRS Business Master File (BMF). A small number of large religious filings not directly connected to health care delivery organizations (the General Conference of Seventh-day Adventists, the Evangelical Lutheran Church in America, the United States Conference of Catholic Bishops, and the Baptist

Convention of Texas) are dropped. We limited the IRS BMF extract to observations containing TINs of hospitals or wholly-owned chain home office TINS.

- Annual SEC 10-K Filings. We used these data to group together hospitals that have a common owner. We limited TIN pairs to those between our list of wholly-owned TINs and hospitals.
- S&P Capital IQ M&A Transactions. We subset the file for consummated M&A transactions involving hospitals and used these data to correct for lags in other data sources.
- Irving Levin Associates Deal Search Online Database and Health Care Services Acquisition Reports. We used consummated hospital mergers and acquisitions in the files to correct for time lags in our other data sources. As with many of the other files, we limited TIN pairs to those between our list of wholly-owned TINs and hospitals.
- PECOS Enrollment Associations File. We used role codes, which describe the relationships between providers and their managing and owning entities, along with providers' organizational structure (LLC, partnership, etc.) to identify ownership relationships in the data.
- Hand coding. There is a handful of ownership relationships we do not observe in the data. We were able to identify some of them when visually screening our list of TIN pairs. We included these relationships as inputs to the network algorithm.
- Networks of Wholly-Owned TINs (from Step 3). Finally, we used these data because we wanted the hospitals to be linked to one of these networks.

In this hospital TIN-TIN dataset, we identified hospital TINs that are linked to two or more distinct networks of wholly-owned TINs (the output from step 3). It is likely that these hospitals have multiple owners (e.g. joint ventures). We excluded these hospitals at this stage to avoid combining two or more networks connected by jointly owned providers (but added them back in at a later stage).

Using R's igraph network library operating on the hospital TIN-TIN dataset and the dataset containing networks comprised of wholly-owned TINs, we connected hospital TINs to the networks of their owning and managing entities. At this stage we revisited the set of hospital TINs initially linked to multiple networks. For hospitals that are jointly owned, we used ownership percentages from the PECOS enrollment association file and online research to assign the hospitals to the network with majority ownership. For jointly owned hospitals that do not have a majority owner (e.g. Centura Health, Duke LifePoint) we created a new network for each set of hospitals with common multiple owners.

Step 4: Identify the owning and managing entities for physician practice organizations and add these practices to networks of owning entities.

We identified ownership relationships for physician practices in many of the sources we used to create our list of wholly-owned TINs and the hospital skeleton including: PECOS enrollment associations file, IRS 990s from Guidestar (now Candid) and AWS, BMF data, Annual 10-K Filings, S&P Capital IQ M&A Transactions, Irving Levin Associates Deal Search Online Database and Health Care Services Acquisition Reports, and an updated list of academic medical groups originally created by Pete Welch and colleagues (Welch and Bindman 2016).

In addition, we conducted analyses in Medicare claims data to identify physician practices billing as hospital outpatient departments. We adapted the algorithm developed by Neprash et al. (JAMA 2015; 175(12):1932-39) for identifying physician practices that are financially integrated with hospitals and operate as hospital outpatient departments (HOPDs). After excluding non-physician NPIs and NPIs that bill predominantly for inpatient care (e.g. anesthesiologists, critical care specialists), we matched physician claims in the Carrier file to claims in the Outpatient file based on a combination of beneficiary ID, service date, procedure code, and/or servicing NPI. Claims were classified as HOP (hospital outpatient) claims if: they have a hospital outpatient department place of service code or if a physician claim in the

carrier file matches a claim in the outpatient file. For each NPI, we computed the percentage of claims delivered in a hospital outpatient setting. For each physician practice TIN containing an NPI billing some portion of their claims as HOP, we computed a measure of how concentrated the NPIs' HOP claims are in a single hospital, or a set of hospitals owned by the same system. We then classified a physician practice TIN as integrated with a hospital based on: 1) % of NPIs in the TIN classified as integrated with a hospital, and 2) the concentration of TIN HOP claims in a hospital or hospital system. For TIN-CCN/ System pairs for which the above criteria are met, we identified physician practice TIN as financially integrated with the CCN with the plurality of matched claims and the physician practice TIN as being a member of the same system as the hospital (CCN).

Analogous to the process outlined in Step 3, we created a dataset of TIN-TIN pairs comprised of a physician practice TIN and the TIN of the entity that owns or manages the practice. Using R's igraph network library operating on the physician practice TIN-TIN dataset and the dataset containing networks comprised of wholly-owned TINs and hospital TINs, we connected physician practice TINs to the networks of their owning and managing entities.

Step 5: Qualify networks as health systems if they meet the definitional requirements.

The final step in identifying health systems was to ensure that the set of providers associated with each network ID meets the minimum criteria for qualifying as a health system.

Inclusion of a short term general acute care hospital

We identified short term general acute care hospitals based on the hospital's primary service type (see step 1 for details). Any network containing at least one short-term general acute care hospitals was deemed to have met this criterion.

Inclusion of at least 10 primary care physicians billing primarily to TINs included in the network

For each primary care physician, we tabulated the number of Medicare and commercial claims billed through each practice TIN. For each network, we then tabulated the number of PCPs that billed a plurality of claims through one of the network's practice TINs. Any network with at least 10 PCPs billing primarily to network practice TINs was deemed to have met this criterion.

Inclusion of at least 50 total physicians billing primarily to TINs included in the network

For each physician we tabulated the number of Medicare and commercial claims billed through each practice TIN. For each network, we tabulated the total number of physicians that billed a plurality of claims through one of the network's practice TINs. Any network with at least 50 physicians billing primarily to network practice TINs was deemed to have met this criterion.

Minimum set of providers located within a single hospital referral region (HRR)

Each physician practice and hospital was assigned to a HRR based on zip code. Any network with the minimum set of providers located within a single HRR was qualified as a health system.

II. Classification of Health System into Categories

For descriptive analyses, we classified each health system into one of five mutually exclusive categories based on size, and the ownership type and academic mission of the system's hospitals. Category assignment was made sequentially in the order shown below.

Assessing hospital ownership and teaching status

The POS and AHA survey data include a variable to describe the hospital's ownership. The types of ownership in each of these databases vary slightly but ownership type is correlated across these databases for the vast majority of hospitals. In cases where there were disagreements across sources, we implemented the following rules:

1. First default to relying on the POS data.
2. If the POS data are missing, "Other," or "Unknown," then use the AHA data.
3. Define hospital ownership as public when the AHA ownership is one of the following values: "CITY", "CITY-COUNTY", "COUNTY", "STATE", or "HOSPITAL DISTRICT OR AUTHORITY."
4. Define hospital ownership as non-profit when the AHA hospital ownership is given as "OTHER NOT-FOR-PROFIT" and the POS hospital ownership is "PRIVATE (FOR PROFIT)." Upon closer inspection, most of these hospitals appeared to be non-profit as opposed to for-profit and were thus reflected more accurately in the AHA survey data.

Assignment of health systems to a category

1. ***Academic Health System***. A health system was categorized as academic if it met either of the following criteria: a) total graduate medical education payments to the system's hospitals exceeds $30,000 per general acute care bed and at least 33% of the system's general acute care beds are in an AHA major teaching hospital; or b) the system met just one of these criteria but received at least $25 million dollars in total graduate medical education payments during the calendar year. Data on hospital graduate medical education payments was obtained from HCRIS. A hospital was classified as a teaching hospital if it met either of the following criteria: a) the hospital is described as a major teaching hospital in the American Hospital Association (AHA) data; or b) the hospital receives at least $10 million in graduate medical education payments.
2. ***Public Health System***. A health system was categorized as public if it met the following criteria: the system is not classified as an academic health system and a plurality of the system's hospital beds are located in publicly owned hospitals.
3. ***Large Not-For-Profit Health System***. A health system was classified as large not-for-profit if it met the following criteria: the system is not classified as academic or public, is comprised of at least 50 primary care physicians located in a single Hospital Referral Region (HRR) and at least 100 primary care physicians across all HRRs, and a plurality of the system's hospital beds are located in not-for-profit hospitals.
4. ***Large For-Profit Health System***. A health system was classified as large for-profit if it met the following criteria: the system is not classified as academic, public, or large not-for-profit, is comprised of at least 50 primary care physicians located in a single Hospital Referral Region (HRR) and at least 100 primary care physicians across all HRRs, and a plurality of the system's hospital beds are located in for-profit hospitals.