CrossMark

RESEARCH ARTICLE

Building an Open Health Data Analytics Platform:
a Case Study Examining Relationships and Trends
in Seniority and Performance in Healthcare Providers

A. Ravishankar Rao[1] (iD) · Daniel Clarke[1] ·
Maryelena Vargas[2]

Abstract With a movement toward an open data-driven approach, governments worldwide are releasing public data in an effort to increase transparency. Despite the availability of this data, many factors make it difficult for anyone to extract knowledge from it. The relationship between seniority and performance is a controversial issue in many fields; using an open and reproducible framework, we investigate the relationship in open healthcare data. Using data from the Center for Medicare and Medicaid Services covering 895,000 practitioners and 3000 hospitals, weak but statistically significant correlations between graduation year, a proxy for seniority, and the hospital value-based performance score were found in 29 of 74 specialties (Spearman rank correlation values < 0.164, p value < 0.05). This result represents 7% of US healthcare practitioners and over 75% of medical practitioners in several specialties. With 5 years of data (2009–2014) from the New York Statewide Planning and Research Cooperative System (SPARCS), we found weak but statistically significant correlations between graduation year and cardiac surgery outcome measures (Spearman rank correlation value -0.096, p value < 0.0005). An unsupervised clustering K-means-based algorithm

✉ A. Ravishankar Rao
    ravirao@fdu.edu

    Daniel Clarke
    danieljbclarke@gmail.com

    Maryelena Vargas
    mvargas1@fdu.edu

[1] Gildart Haase School of Computer Sciences and Engineering, Fairleigh Dickinson University, Teaneck, NJ, USA

[2] The Henry P. Becton School of Nursing & Allied Health, Fairleigh Dickinson University, Teaneck, NJ, USA

for finding outliers was also applied to these datasets. It captured a unique trend in the number of nurse practitioners which was increasing rapidly since 2010. It also revealed consistencies in practitioner placement throughout hospitals. Our findings suggest that the training of healthcare professionals appears to be robust and positions them for long-lasting and consistent careers across the majority of specialties.

Keywords Hospitals · Information dissemination · Medicare · Surgeons · Nurse practitioners

1 Introduction and Motivation

It is paramount to quantify and measure health outcome as this leads to improved care delivery, cost reduction, and efficiency [1]. This has led to large-scale data-driven approaches in health at both the individual and national levels [2]. Individuals use a wide variety of personal monitoring devices such as fitness trackers to measure variables that impact health [3]. Many government institutions worldwide are moving toward open data and transparency. The Center for Medicare and Medicaid Services (CMS) in the USA releases data continually about the performance of hospitals nationwide (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4]. Pharmaceutical companies release data to the public, through initiatives such as the Yale University YODA project [5]. Rao et al. review benefits provided by this open data movement for healthcare [6].

Such data-driven and transparency approaches to medical costs have dramatically reduced medical spending in the California Public Employees' Retirement System [7]. The article [7] observes, "It requires patients to have ready access to comprehensible price and quality information. Such transparency is not commonplace. Even when this information is available, consumers with cognitive impairments, or who are overwhelmed with illness and other demands would have trouble making the best use of it."

This paper focuses on publicly available data as large amounts of public data already exist, and value remains to be extracted therefrom. Leveraging publicly available data allows analyses to be easily replicated and expanded by researchers worldwide. This should spur innovation [8] in the field of healthcare data analytics. Johnson [9] indicates that in addition to making data available, it is also important to run code consistently, thereby enabling end-to-end reproducibility of analysis. Johnson [9] and Paul [8] recommend that making code open source is a promising direction to ameliorate the lack of reproducibility in scientific research, where more than 70% of researchers have been unable to reproduce another scientist's experiments [10]. The code used in this paper is available freely through an open source repository on Github [11].

A major challenge in extracting value from existing data lies in being able to organize and integrate the numerous datasets scattered across multiple federal and state level agencies in the USA. CMS provides data about healthcare providers and hospital ratings (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4], the Center for Disease Control has data about disease incidence [12], the New York State SPARCS initiative [13] has data about patient discharges in NY State, and the New York State Department of Education [14] has information about graduation records for providers. The current paper describes a viable approach to

combine information from multiple datasets, which is utilized to answer specific questions about the relationship between seniority of practitioners and performance, and understands trends in graduation rates. This research is also timely as there are several organizations including the NIH [15, 16], the Robert Wood Johnson Foundation [17], and the Russell Sage Foundation [18] that are offering grants to support the analysis of existing healthcare datasets and novel methods to analyze and interpret such data. Clearly, we are awash in data.

In a systematic review, Fung et al. [19] observed that a lack of transparent, explicit, data-driven performance measurement and feedback mechanisms for healthcare providers contributes to deficiencies in the delivery of effective safe and patient-centered care. The Institute of Medicine (IOM) has also identified these same areas as essential for a well-functioning healthcare system. The data continuously released by the CMS (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4] is a step in the right direction, as it includes measures of hospital performance.

The availability of public healthcare data provides opportunities to tackle a wide range of interesting and relevant questions, such as the relationship between physician spending and outcomes [20], the effect of price transparency on consumer health choices [21], or the nature of the workforce in the healthcare industry [22]. Since these are broad research areas in themselves, we restrict the scope of the current paper to important questions concerning the relationship between age and performance and trends in graduation across medical specialties. There are professions with mandatory retirement ages. The Federal Aviation Administration has mandated a retirement age of 65 for air traffic controllers and pilots (https://www.faa.gov/other_visit/ aviation_industry/airline_operators/airline_safety/info/all_infos/media/age65_qa.pdf). One of the grounds for this rule is that physical and mental skill may decline significantly after age 65. It has been difficult to create such a rule for healthcare professionals as the link between age and performance has not been well established.

The application of big data analytics [2] to open healthcare data provides a novel opportunity for exploring these relationships between seniority and performance in healthcare practitioners. This is important as it could provide comparative evidence of the correlation between seniority and performance through public reporting. This work can also have an impact on patient safety and patient centeredness, particularly about individual providers and practices. Analyzing and releasing performance data may stimulate quality improvement activity at the hospital level particularly among surgical specialties. This contributes to transparency of the healthcare system and informs consumers of healthcare. Whether provider age is associated with patient outcomes remains largely unknown. It is possible for providers' skills to become outdated with advances in scientific knowledge and technology and changes in clinical guidelines.

The relationship between the age of medical professionals and the quality of care and outcomes is controversial and is receiving increasing attention. On the one hand, researchers such as Choudhry et al. [23] observed an inverse relationship between the number of years that a provider has practiced and the quality of care provided. They reported that older physicians have decreased clinical knowledge, adhere less often to standards of appropriate treatment, and perform worse on process measures of quality with respect to diagnose, screening, and preventive care. They suggested methods to overcome this problem such as continuing professional medical education. Neumayer et al. [24] also showed age-related decline in surgical performance. Jena et al. [25]

showed that mortality rates for cardiac patients decreased when senior cardiologists were away for the National Cardiology Meeting in the USA. Schenarts [26] surveyed a growing literature showing that more experienced physicians and surgeons produce paradoxically worse clinical outcomes. Anderson et al. [27] showed that outcomes produced by very senior surgeons (60–80 years old) were associated with higher morbidity and mortality.

On the other hand, we could expect more experienced practitioners to provide better care. McAlister [28] did not find a negative association between physician experience and outcomes. This may depend on the specialty. Li et al. [29] determined that patients treated by junior and intermediate emergency physicians in emergency departments had higher mortality as compared with senior emergency physicians. A total of 59 physicians were studied.

Other observations have shown that the seniority of providers does not affect patient outcomes. Cunningham et al. [30] measured patient outcomes delivered by three categories of surgeons based on differing levels of seniority. They tracked 278 patients found no statistically significant difference in patient outcomes. An analysis by Waljee et al. [31] showed that for most surgical procedures, surgeon age did not predict operative risk.

The use of data at the national level can address issues of small sample size, disease, or specialty scope. This provides a novel approach to investigate whether physician age is associated with patient outcomes.

Tsugawa et al. [32] conducted a study which showed that the adjusted 30-day mortality rate of patients increased with the age of the physician, with a sharp rise in mortality when the age of the provider was greater than 60 years of age. The findings of Tsuguwa et al. [32] have policy implications regarding the importance of patient outcomes as one component of an assessment of how provider practices change over a career. Melville [33] notes that though age-based assessments of provider competency is a very controversial subject, an ever increasing number of providers continue working well into older age. Age-related changes including declines in visuospatial skill or manual dexterity may have an impact on their ability to practice, particularly in surgical specialties.

Efforts at mandatory self-assessment in medicine have not fared well. Some hospitals such as Stanford Hospital and Clinics require physicians older than 75 years of age to undergo peer evaluation of clinical performance and other testing every 2 years. A key argument favoring age-related assessment is that those with cognitive decline or mild cognitive impairment may fail to recognize or be willing to act on their own decline as shown by Davis et al. [34]. This suggests that physicians have a limited ability to accurately self assess. The processes currently used to undertake professional development and evaluate competence may need to focus more on external assessment, which highlights the importance of applying big data analytics to examine the correlation between seniority and performance since it impacts patient outcomes.

Kupfer [35] showed that many studies are limited by small sample sizes. This suggests that a national big-data approach has the potential to overcome such limitations. The presence of such seemingly conflicting studies indicates evidence of the need to synthesize significant findings from research in a user friendly and readily available manner that can be used by providers to provide higher quality of care and achieve better outcomes for patients.

In summary, the presence of age-related cognitive decline in humans [36] suggests that performance scores should decline with increasing seniority. This is countered by the observation that experienced practitioners could provide better care (McAlister [28], Li et al. [29]).

Hence, a specific research question addressed in this paper is the controversy in the medical field about the relationship between age and seniority level of providers performing different procedures on the patients and outcomes [31, 33, 35, 36].

A related question concerns the patterns among the seniority levels of different providers. What are the similarities and differences in graduation trends among different medical disciplines? This question is important because it provides evidence that helps guide capacity planning for providers required in different specialty areas. For instance, Meagher [37] published a paper in 2016 that reviews data compiled by the American Hospital Association in 2011, and showed that general surgery is one of the specialties predicted to have the largest projected shortfalls by 2020. It would be interesting to find other specialties similar to general surgery that have similar shortfalls, and to be able to identify them through automatic clustering techniques. Furthermore, we can see that it takes several years for such data to be compiled and analyzed, as there was a lag of 5 years in the case of Meagher's study [37]. The availability of public healthcare data offers a faster route to perform similar analysis, to perform it with less time lag, and automatically re-run the analysis on a quarterly basis, as new data are released every quarter by the CMS office (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4].

Once we gather the graduation years of providers (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4], we can leverage this information to investigate additional questions. For instance, understanding which healthcare specialties are similar in terms of graduation trends can help educational institutions better prepare for serving fast-growing specialties such as nursing, nurse practitioners (NPs), and physical therapists. Groups of similar specialties can be identified through automatic cluster analysis techniques, and such groups can be utilized for curricular integration, which is an important strategy that educational institutions can deploy [38].

Accordingly, we formulate the following research objectives:

1. To investigate the relationship between seniority and performance in the CMS dataset (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4]
2. To investigate the relationship between cardiac surgery outcomes and surgeon seniority in the New York State SPARCS dataset [10]
3. To determine patterns and trends in the graduation of practitioners across all the specialties in the CMS dataset

2 Related Work

We present a framework that can analyze large public healthcare datasets released by CMS [4] and New York State SPARCS [10, 39]. We integrate multiple data sources from disparate government agencies at the federal and state levels in order to answer

specific research questions. In the current paper, we demonstrate the detailed use of four different data sources, though our approach and framework can be generalized to multiple sources. We use different techniques including outlier detection to identify problems in the public healthcare datasets. For instance, the graduation years of a few providers in the CMS (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4] data were well beyond the life expectancy of current providers.

Bullinger et al. [40] support an open innovation ecosystem for healthcare, where contributions can be made by multiple segments of society, including patients, care givers, healthcare providers/practitioners, family members, and the interested public. Patients and the interested public need to understand how government healthcare expenditure impacts public health and can play an important role by being vigilant about available options. Important questions include the following: Where is the money being spent? What do procedures cost? Where are the inefficiencies in the system? Is improvement in patient outcome commensurate with expenditure? Where can a patient go to find affordable care for a specific condition? Are all practitioners providing the same level of quality of care? There are many challenges to overcome before such questions can be tackled. Healthcare systems are usually proprietary, siloed, and have little interoperability [15]. Open systems and architectures would lead to cooperation, improvement of interoperability, transparency, and increased innovation among healthcare players [41]. Safran et al. [42] recognize the importance of educating the public in the use of aggregated personal health information. For instance, Rosenkratz [43] studied the availability and accuracy of price and performance information of radiology practices on public websites, and found that performance information usually presented is not of adequate clinical quality. An increasing number of researchers are applying machine learning and advanced analytic techniques to interpret healthcare data [44, 45].

There is a high degree of consensus among several organizations, such as the American Association of Colleges of Nursing (AACN) and the Institute of Medicine (IOM) on enhancing interdisciplinary partnerships that can help healthcare providers and organizations to build and operationalize "a culture of health." This will require substantive shifts in current approaches and delivery systems beyond the patient and provider. The IOM report [46] calls for nurses to partner with physicians and other healthcare professionals to redesign healthcare in the USA. Cost-effective coordinated care is already provided in multiple settings by NPs [47, 48]. Yet, these contributions are not fully recognized by policymakers and government, which continue to subject NPs to outdated laws requiring a collaborating physician. The state of New Jersey utilizes this outdated regulation to require that NPs practice with a joint protocol with a collaborating physician, N.J.A.C 13:37-6.3 (NJ Board of Nursing [49]). Such laws limit effective workforce planning. Better workforce planning management and an improved information infrastructure are necessary for collecting and analyzing workforce data for the US healthcare system to prepare for the provision of primary care for the 72 million Americans that will be 65 and older by 2030 [68] and the 30–34 million newly insured Americans after the inception of the Affordable Care Act (ACA).

Systematic assessment and projection of workforce requirements by role, skill mix, region, and demographics are required to inform changes in practice, policy, and education. The 2010 ACA [50, 51] supports research that compares the clinical

effectiveness of health treatments, the creation of a national healthcare workforce, and systematic monitoring of the supply of healthcare workers across professions. Since the new US administration in 2017 calls for a repeal and replacement of the ACA, more than ever a review of available healthcare data can help develop methods to predict workforce needs and inform the most effective method to replace or complement the ACA. An example is the re-structuring of the ACA with a system such as Medicare-for-all that partners NPs with other healthcare professionals to coordinate and provide primary care. This requires coordinated data collection on the healthcare workforce at the state and regional levels. Data collected must be timely and publicly available to ensure a healthcare system which provides seamless, affordable, quality care that is accessible to all.

This creates a unique opportunity to utilize a data integration model to identify and report on systematic variation in aggregate outcomes. The Kaiser Foundation [22] calls for the use of data-intensive techniques to fully understand the impact of current healthcare laws in the USA on vulnerable populations.

The techniques provided in our current paper provide a methodology to make the analysis of healthcare data accessible to a wider pool of researchers, and dovetails with the principles embodied in the 2010 ACA. Data can be collected and analyzed from multiple perspectives, including cost metrics, timeliness of service, and outcomes. The free availability of data analytic platforms like the one in this paper [8] will facilitate the exploration and interpretation of healthcare datasets by patient advocacy groups and concerned citizens.

3 System Design and Approach

Our original data analysis framework [4], shown in Fig. 1 consisted of the following steps: data cleansing/ETL, data joining, feature engineering, clustering classification and prediction, visualization of results, interpretation, and reporting.

The framework builds on existing open source Python components, including Pandas, SciPy, Scikit-Learn, and Matplotlib [52]. Python Pandas is popular for big-data analytics in finance [53], but not in the healthcare industry. Python Pandas is advantageous as it integrates tabular and statistical modeling while interfacing with a visualization and plotting package via Matplotlib. The current research shows that Python Pandas is readily applicable to healthcare datasets.

The SciPy toolkit, popular in physics [54] and finance [55], permits rapid exploration of datasets and algorithms. The Scikit-Learn Python library [56] provides advanced machine-learning capabilities including clustering, classification, and prediction. Matplotlib and Graphviz provide visualization capabilities. Figure 2 contains a schematic of the different Python modules used.

To improve querying speeds, the cleaned datasets were migrated from CSV files to a MySQL database. This facilitates database operations, given the size of the datasets, which is in the range of several gigabytes. It also provides a transparent way to access the data sets in the scripts. The reader can view sample code to perform these operations in our open source repository on Github [11].

The key innovation in our research is to present an end-to-end solution that can start with the raw data sources, join them, run the analytics procedures, and answers
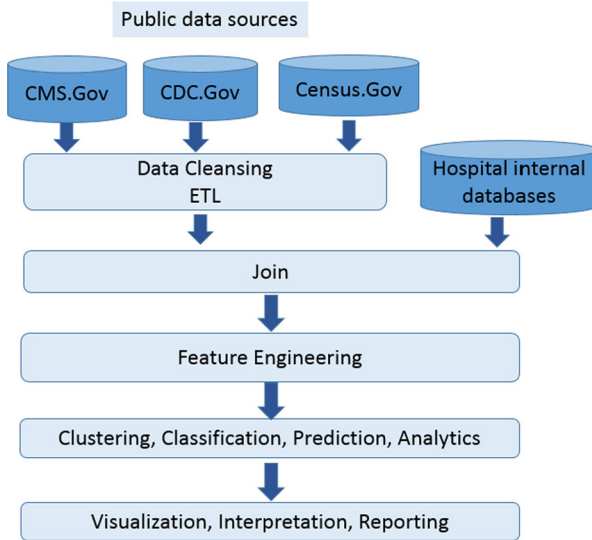
Fig. 1 Data Analysis Framework

interesting research questions. Our solution is based on open-source software, and the entire analytics pipeline has been tested with public healthcare datasets, and made available freely. This makes our work replicable by other researchers, who can also modify and enhance our existing analytics pipeline. An important feature of our solution involves the use of Python-based Jupyter Notebooks [54], which has been proposed as a way to overcome the lack of reproducibility in medical studies [9]. Our code has been released in the form of Jupyter Notebooks, which allows other researchers to view the code and reproduce the results in the current paper within a single software environment.
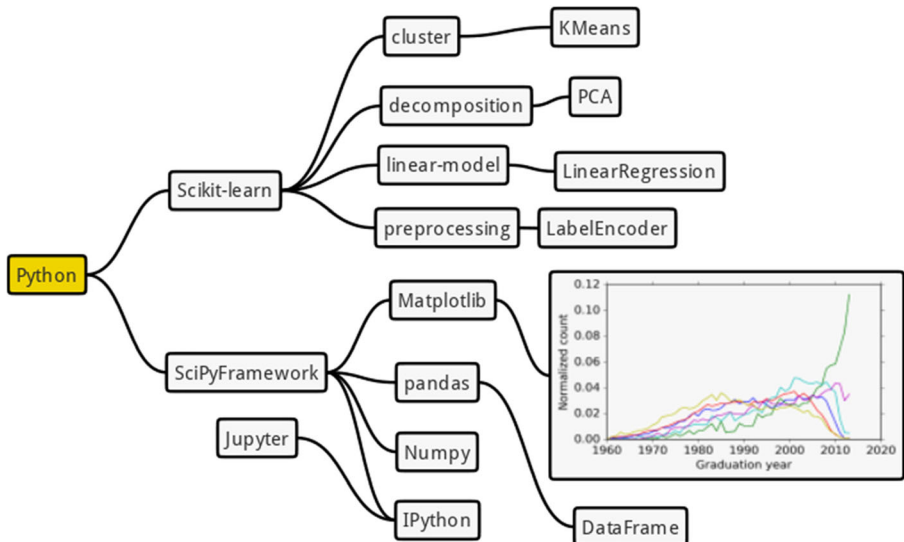

Fig. 2 The Python ecosystem used to perform the analytics in the current paper

4 Methods

Two sources of public health data were used, consisting of CMS data and New York SPARCS data. Since the CMS data does not contain information about outcomes delivered by individual practitioners, New York SPARCS data was used, which contains cardiac surgery outcomes by surgeon. This data serves to illustrate the recommended analytic procedure. The current paper extends previously reported results in [45] .

4.1 CMS Data

CMS data (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4] was processed using the pipeline in Fig. 1. Several errors in the CMS data were identified including data entry errors [57]. These rows were eliminated, and we verified that this elimination does not adversely affect our analysis and conclusion.

CMS data that resides in the following files were joined.

1. A file containing the hospital id and physician information: https://data.medicare. gov/Physician-Compare/National-Downloadable-File/s63f-csi6, henceforth referred to as "National-Downloadable-Physician-Compare". Sample data is shown in Figure 12 (in Supplementary Material).
2. A file containing the hospital id and hospital performance metrics which includes the weighted clinical process of care domain score, weighted patient experience of care domain score, weighted outcome domain score, and total performance score http://www.medicare.gov/hospitalcompare/data/total-performance-scores.html. Sample data is shown in Figure 13 (in Supplementary Material). The description and interpretation of these fields is presented within this website.

The CMS office uses a standardized procedure (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4] to evaluate hospitals across the USA, based on the outcomes for Medicare/Medicaid patients. There are four components or domains that yield the final "Total Performance Score" and consist of the following: (1) clinical care domain comprised of process and outcomes subdomains, (2) patient- and caregiver-centered experience of care/care coordination domain, (3) safety domain, and (4) efficiency and cost reduction domain.

• The clinical care domain is comprised of clinical process measures and measured mortality.
• The patient-and caregiver centered experience of care/care coordination domain is derived from patient surveys.
• The safety domain contains a patient safety measure and measures of infections
• The efficiency and cost reduction domain contains a measure of the Medicare spending per beneficiary

These measures are weighted according to a formula devised by the CMS office. Though these measures are authoritative, in that, they are released by the federal

government, there is an effort to constantly improve them. For instance, the information derived from patient surveys could be biased. It is likely that patients with the best or worst experiences may be more likely to complete these surveys. Some of the survey questions are prefixed with "Always," which may make it hard for patients to register an appropriate response. The creation and validity of these performance measures is a subject of active research and debate [58] and falls outside the scope of the current paper.

Press [59] reviews several factors that need to be taken into account in defining performance measures, including process measures that occur prior to and during the care, applying appropriate risk adjustments. This is followed by patient satisfaction measures and outcome measures including the occurrence of adverse events. The relationships between different metrics continues to be explored. For instance, Kernisan [60] determined that there was no association between a hospital's self-reported safe practices survey score and risk-adjusted patient mortality. Leonardi [61] compared different websites that contained hospital performance data, and found the CMS Hospital Compare website to be the most transparent in terms of explaining data sources and statistical methods that allow their exact results to be duplicated.

Consequently, we choose to use "as-is" the federally approved measures in the CMS database. Data from 895,431 unique practitioners in Figure 12 was joined with data from 3089 unique hospitals in Figure 13. To this joined data, correlation and regression analysis was applied, available in Python Scikit.

4.2 Data from New York Sparcs and New York State Education Department

The NY State SPARCs program contains detailed information about cardiac surgery outcomes, including the physician ID for cardiac surgeons and the risk-adjusted outcome for their patients. The New York State Education department website [14] contains data about the physician ID, specialty, and year of graduation, which is a proxy for experience. Note that the Education department's graduation year information is not downloadable directly. A custom Python script was written to query the website and create a repository of graduation year information. Sample data are shown in Figures 14 and 15 (in Supplementary Material). Correlation and regression analysis was applied, available in Python Scikit.

The K-means clustering algorithm [56], available in Python Scikit-learn was applied to understand the inherent groupings in the different specialties according to graduation year distributions. K-means clustering is an unsupervised learning technique which automatically identifies clusters in data for a given number of clusters, denoted by the variable K. The value $K = 6$ was used to find clusters among various specialties. The K-means clustering algorithm operates on an ensemble of objects, where each object is described by a feature vector. For instance, an object for the specialty "NURSE PRACTITIONER" is described by a feature vector comprising 54 dimensions, where each dimension consists of the number of nurse practitioners that graduated in a given year ranging from 1960 to 2014. In other words, the dimension represents a bin with a width of 1 year. So the 54th dimension of the feature vector is the number of nurse practitioners that graduated in 2014, as determined by the "Graduation Year" field in the CMS database. The feature vector is normalized so that the sum of all the normalized feature components equals 1.

The *K*-means clustering algorithm iteratively groups the feature vectors according to similarity, so after a given number of iterations, we arrive at a grouping of a pre-specified number of clusters, say $K = 6$. Further details about the algorithm can be found in [62].

The choice of $K = 6$ is based on an expectation of interpretability. We could have chosen K to be very small, say 2 or 3, but that would lump together groups that do not have much in common. On the other hand, if K is very large, say greater than 20, we would have many small clusters, which would be more difficult to interpret. We use $K = 6$ as it offers a reasonable tradeoff between generality and specificity. In our code, K is a parameter, and researchers are able to experiment with different values of K to better understand the groups in the data. Due to space restrictions, we have chosen to focus on the value of $K = 6$. Our interpretation is not critically dependent on the choice of this parameter.

Similarly, our choice of 54 dimensions for the feature vector to describe a specialty is based on the graduation year being a discrete number from 1960 to 2014, consisting of 54 values. We could have binned the graduation years differently, say choosing 27 dimensions where two consecutive years, say 1960 and 1961 are mapped to the first bin, the years 1962 and 1963 are mapped to the second bin and so on. We have chosen to conduct the analysis at the finest level of resolution available in the data, which is 1 year, resulting in 54 dimensions.

5 Results

Results from the analysis of two separate sources, CMS data and New York State SPARCS are presented.

5.1 CMS Data

Table 1 compares the number of medical practitioners in the current CMS dataset to the actual number of medical practitioners determined by AAMC [63]. There is remarkably high coverage in the CMS dataset of the total number of practitioners in USA, and exceeds 75% for several specialties.

A heatmap is used to explore the relationship between the graduation year and total performance score for each specialty. This provides a rich description of the variation between these two dimensions, and helps interpret the correlation coefficient. Figure 16 illustrates the computation and visualization of the heatmaps. The CMS data is projected onto a 2-dimensional space with the axes consisting of graduation year and total performance score. The *x* axis, representing the graduation year is divided into discrete bins, with one bin for each graduation year, say 1960 or 1961. The *y* axis, representing the total performance score is divided into discrete bins, say a bin for total performance score between 40 and 40.5, and the next bin for total performance score between 40.5 and 41, and so on. The number of occurrences of data points within each 2-dimensional bin is counted. The count in this 2-dimensional bin is colorized and represented as a 2-D plot known as a heatmap, as shown in Fig. 3.

Figure 3 shows interesting trends in the number of graduates from years 2006 to 2014 which are further examined in Fig. 4 via projections of the heatmaps. The years

Table 1 The column "Medicare count" shows the number of unique practitioners in the CMS. The column "AAMC Count" shows the number of medical practitioners by specialty as determined by the American Association of Medical Colleges in 2013 [Reference [63]]. The column "Percentage covered" is the percentage of the AAMC count covered by the current Medicare dataset for the given medical specialties. A few chosen specialties show that the CMS dataset is representative of more than 75% of all medical graduates for several specialties

Specialty	Medicare count	AAMC count	Percentage covered
Internal medicine	96,975	111,047	87.33%
Family practice	81,901	108,917	75.20%
Anesthesiology	35,667	40,758	87.51%
Obstetrics/gyn.	31,399	40,826	76.91%

are shown in different colors. Consider a ratio L for a given year for a given specialty defined as follows:

$$L = \frac{N1}{N} \tag{1}$$

where $N1$ is the number of practitioners for the given year and specialty who practice in a hospital with total performance score lower than the mean total performance score of 39.6, and N is the total number of practitioners for the given year and specialty. For each year and specialty, we get a different value of L, which is shown in the inset in each subplot. A value of L close to 50% indicates that the practitioners for a given year and specialty are spread evenly across hospitals based on their performance score. If L is much higher than 50%, it indicates that practitioners are working in hospitals with lower performance scores.

Figure 5 shows the results of applying linear regression to data that includes the year of graduation for all medical practitioners versus different hospital metrics provided by CMS. For instance, the blue dots and the blue line represent the correlation and fit between the graduation year for a provider and his or her affiliated hospital's total performance. Similarly, the green and red dots and lines show the correlation and fit for other hospital metrics, such as the "weighted clinical process of care domain score" and "total performance score." Detailed descriptions of these metrics are provided in [4].

Table 2 shows the result of computing the Spearman's rank correlation coefficient between the graduation years and total performance score across various specialties. Spearman's rank correlation was used as the graduation years were discrete and non-continuous variables.

The results of applying the K-means clustering algorithm [28] are shown in Figs. 6, 7, and 8. The feature vector for each specialty contains the normalized counts of practitioners for each graduation year. The K-means clustering is an unsupervised clustering algorithm, and does not need any training data or specific hints about the types of clusters to be identified. The only input required is the expected number of clusters, which was 6. The K-means clustering algorithm identified several interesting trends which are interpreted in the Discussion section.
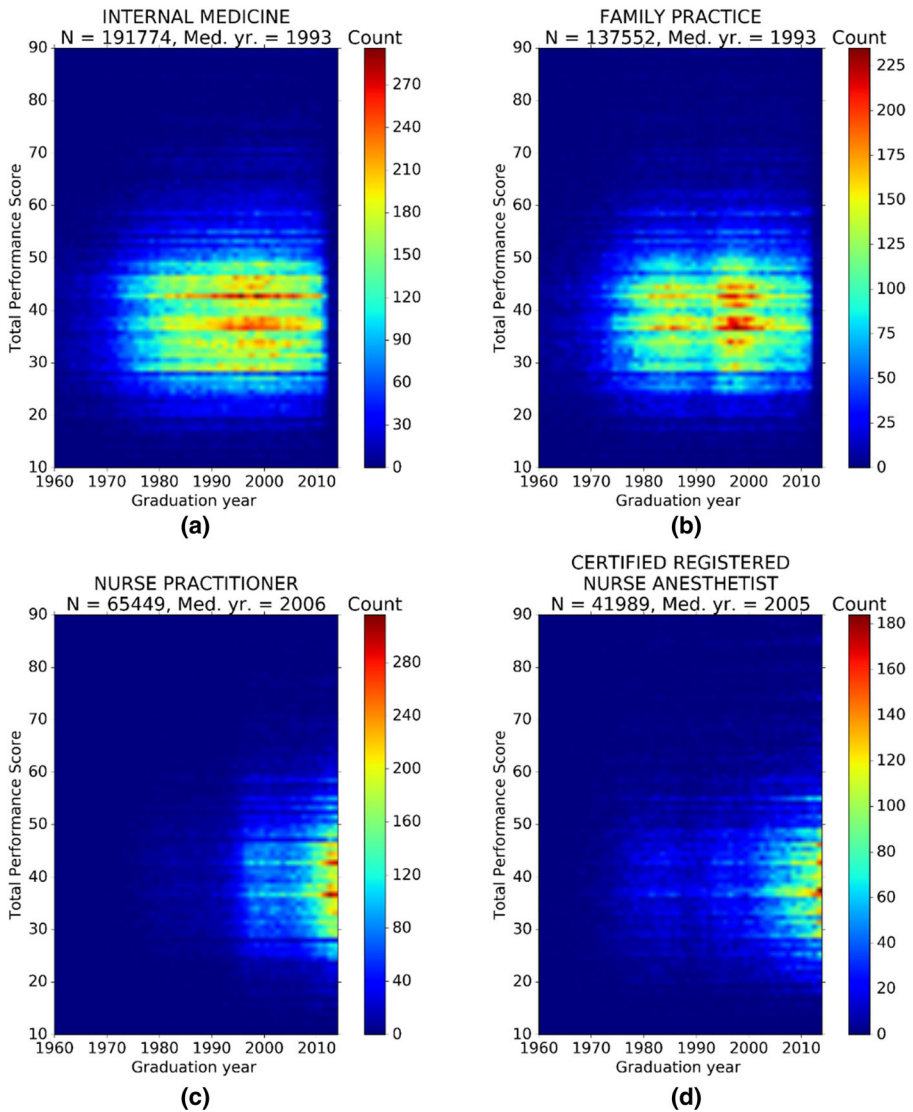
Fig. 3 A heatmap showing the relation between graduation year and total performance score. This is for the specialties consisting of "Internal Medicine," "Family Practice," "Nurse Practitioner," and "Certified Registered Nurse Anesthetist." Practitioners across all hospitals were used. In subplot (a), $N = 191,774$ refers to the total number of relationships between internal medicine practitioners and hospitals. Similarly, "Med. Yr" represents the median graduation year for all internal medicine practitioners across all hospitals, which is 1993. The colors in the map represent the counts at specific (x,y) locations corresponding to combinations of (Graduation year, total performance score)

5.2 Outcomes for Cardiac Surgery from NY SPARCS Data

A detailed list of patient records released by NY SPARCS shows the risk-adjusted patient outcomes for cardiac surgery. We were able to triangulate information from another source, the NY State Education department to retrieve information about
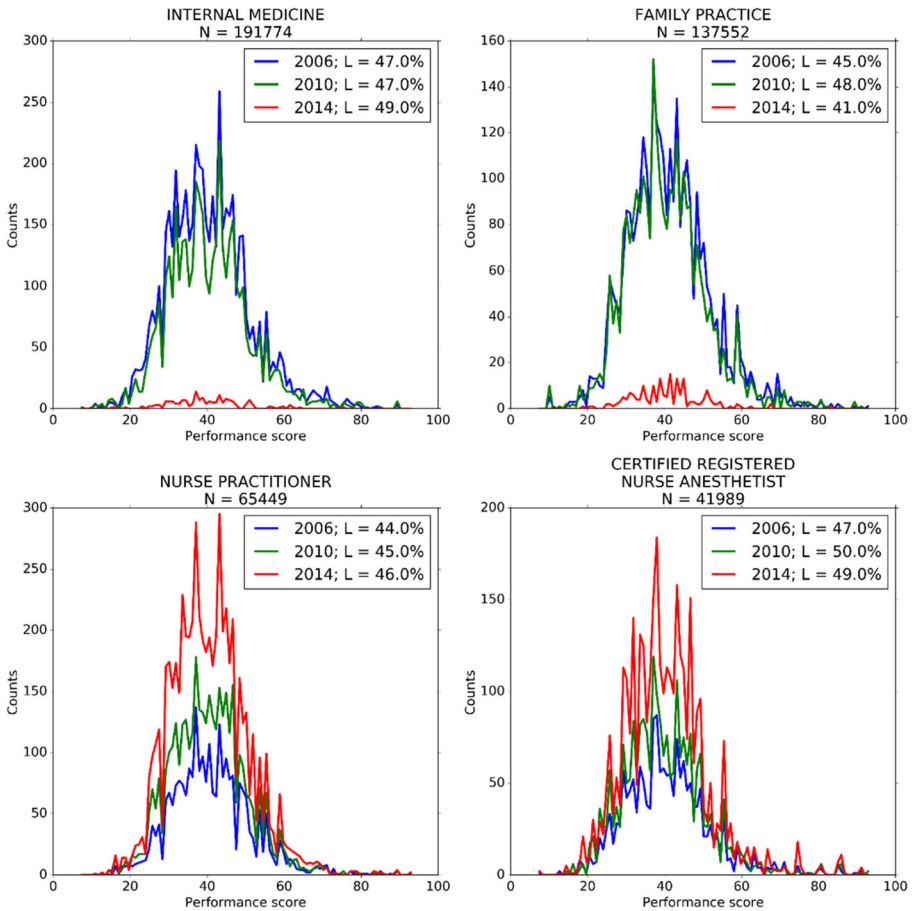
Fig. 4 Trends for recent graduates in the following four specialties: internal medicine, family practice, nurse practitioner and certified registered nurse anesthetist. The inset in each subplot shows the value of L calculated according to Eq. 1. There are three calculated values of L, one for each curve shown for the years 2006, 2010, and 2014. The curves are color-coded in blue, green, and red respectively. For instance, the red curve for internal medicine practitioners shows that far fewer graduates in 2014 practiced in Medicare-serving hospitals as compared to the year 2006. The distribution of these graduates was quite even across all hospitals, with values of L being 49 and 47% respectively. This is similar for family practice. Compared with the red curve for nurse practitioners shows that far more graduates in 2014 practiced in Medicare-serving hospitals as compared to the year 2006. The distribution of these graduates may pick up on a trend of NPs filling the gap left open by internal medicine and family practice quite even across all hospitals, with values of L being 46 and 44% respectively. Further research in this area utilizing large datasets may have policy, educational, and healthcare implications in the USA

graduation years for cardiac surgeons. A scatter plot of this data is shown in Fig. 9. The Spearman's rank correlation coefficient between risk adjusted mortality rate and graduation year was -0.096 (p value < 0.0005). This shows that the correlation is very small.

2. Discussion

Table 2 shows that the Spearman's rank correlation coefficient is statistically significant but small for several specialties. Hence, our result indicates that there is little correlation between the assumed age of a practitioner at a hospital and the
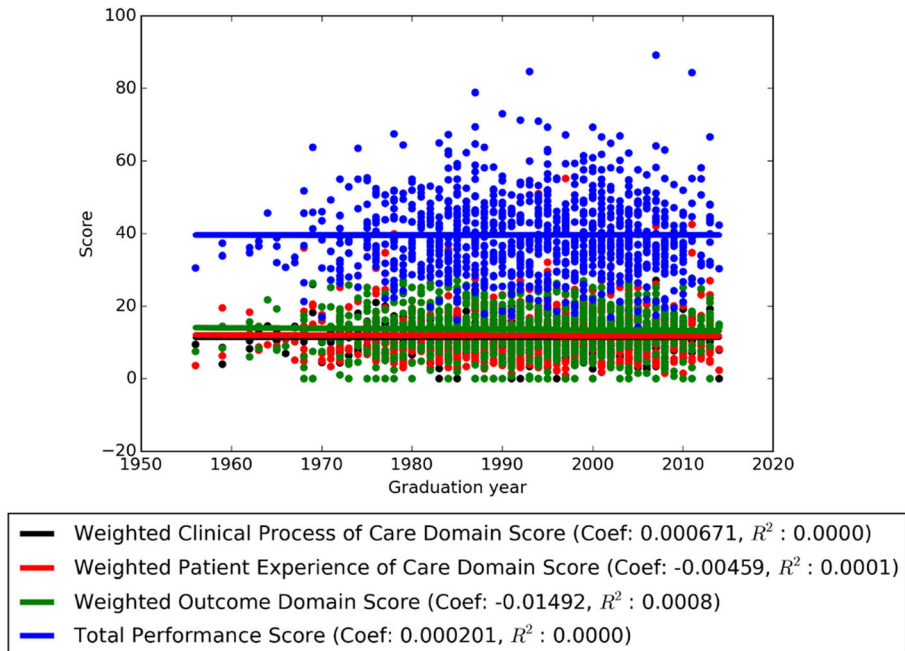
Fig. 5 Variation of the different hospital scores used by CMS versus the graduation year of all practitioners

performance score of that hospital. There are several issues regarding the estimation of the relationship between the graduation year and hospital performance. It is possible that the effects of cognitive decline and improved experience may balance each other out in larger populations of practitioners. Even if age may affect individual practitioner performance, this is not reflected at the institutional level. This may because there is a variety of ages and experience levels of practitioners at each hospital, which confounds a simple model. Furthermore, the Hospital Total Performance score is an indirect measure, and already averages the performance of several individual practitioners and the hospital administration. The Hospital Total Performance score may also depend on the patient pool, with poorer populations likely to experience worse outcomes overall.

The median graduation year for internal medicine practitioners, computed from Fig. 3 is 1994. Assuming that the age of a practitioner is 27 upon graduation, this translates to an estimated median age of 49 for internal medicine practitioners. This agrees with estimates provided by the AAMC [63]. It can be argued that ultimately what matters to a hospital is its total performance score, and it is important to consider all factors impacting this score, including the seniority of practitioners. The analysis in this paper indicates that seniority is not a significant factor. This is especially striking when this finding is contrasted with the relationship among seniority and job performance in other sectors of the economy. The workforce in more successful technology companies is significantly younger than the workforce in less successful companies [64, 65]. The median age of Google employees is estimated to be 29 whereas the median age at Hewlett Packard is estimated to be 41 [64]. The Bureau of Labor Statistics estimates the overall median age of American workers is 42.3 years old. It appears that in the

Table 2 The Spearman's rank correlation between the graduation years for each specialty and the total hospital rating provided by CMS. The values were sorted from the highest Spearman's rank correlation coefficient to the lowest. These correlations were computed across all hospitals. Only specialties whose Spearman's rank correlation had a *p* value of less than 0.05 are shown. *N* is the number of practitioners

Specialty	Spearman's correlation coefficient	Spearman *p* value	R	R^2	Slope	Mean	Standard deviation	N
Geriatric psychiatry	0.164	0.04507	0.22	0.049	0.164	38.638	8.254	150
Anesthesiology assistant	0.11	0	0.075	0.006	0.075	37.643	9.151	1742
Hematology	0.077	0.00781	0.066	0.004	0.055	39.558	9.811	1191
Allergy/immunology	0.063	0.00087	0.053	0.003	0.041	39.625	9.368	2820
Medical oncology	0.058	0.00001	0.053	0.003	0.049	39.724	9.979	5864
Pulmonary disease	0.051	0	0.051	0.003	0.05	39.434	10.317	21,587
Critical care (intensivists)	0.051	0.00025	0.052	0.003	0.055	39.327	10.089	5196
Dermatology	0.047	0.0001	0.054	0.003	0.04	39.928	9.547	6977
Gynecological oncology	0.047	0.02151	0.05	0.002	0.048	39.413	10.198	2369
Physical medicine and rehabilitation	0.045	0	0.051	0.003	0.054	39.812	10.505	12,211
Radiation oncology	0.041	0.00033	0.037	0.001	0.033	39.719	9.969	7714
Thoracic surgery	0.041	0.00506	0.048	0.002	0.047	39.208	10.14	4618
Colorectal surgery (proctology)	0.039	0.0435	0.04	0.002	0.037	38.913	9.963	2709
Urology	0.038	0	0.044	0.002	0.04	39.719	10.538	22,702
General practice	0.037	0.00303	0.041	0.002	0.034	38.898	10.161	6306
Pathology	0.033	0	0.033	0.001	0.032	40.321	10.993	21,236
Gastroenterology	0.029	0	0.031	0.001	0.029	39.3	10.143	27,131
Psychiatry	0.028	0.00039	0.03	0.001	0.023	38.662	9.593	16,178
Internal medicine	0.026	0	0.033	0.001	0.03	39.441	10.221	191,774

Table 2 (continued)

| Specialty	Spearman's correlation coefficient	Spearman p value	R	R^2	Slope	Mean	Standard deviation	N
Otolaryngology	0.025	0.00205	0.026	0.001	0.024	39.55	10.532	15,618
Cardiovascular disease (cardiology)	0.024	0	0.027	0.001	0.026	39.527	10.453	64,306
General surgery	0.024	0.00001	0.031	0.001	0.028	39.437	10.671	34,089
Neurology	0.023	0.00017	0.025	0.001	0.021	39.438	10.033	27,499
Hematology/oncology	0.019	0.00934	0.024	0.001	0.022	39.548	10.1	18,420
Family practice	0.018	0	0.02	0	0.019	39.89	10.205	137,552
Anesthesiology	0.015	0.00167	0.018	0	0.018	39.634	11.082	47,000
Obstetrics/gynecology	0.015	0.00175	0.02	0	0.017	39.121	9.809	43,366
Nurse practitioner	− 0.019	0	0.017	0	− 0.022	40.335	10.176	65,449
Certified registered nurse Anesthetist	− 0.033	0	0.032	0.001	− 0.032	40.102	11.064	41,989
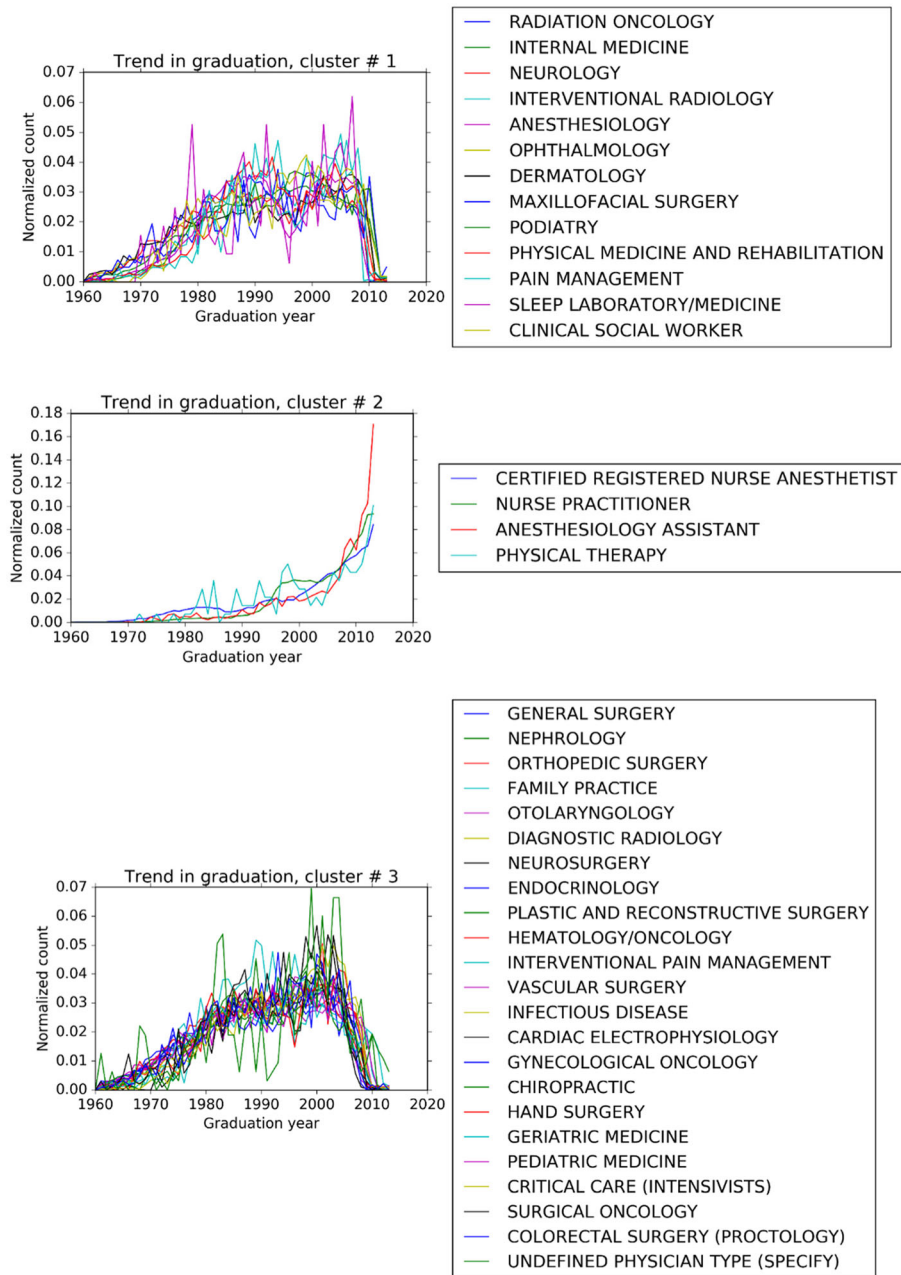
Fig. 6 The results of applying the *K*-means cluster analysis on the graduation data. The *y* axis depicts the normalized count, which is proportional to the number of graduates for a given specialty for a given year. This normalization facilitates the comparison of multiple curves

technology sector, seniority of employees is negatively correlated with the total performance of the companies they work at. It has been estimated that meaningful insights can be derived from data covering 3% of the American work force [64]. The Kaiser
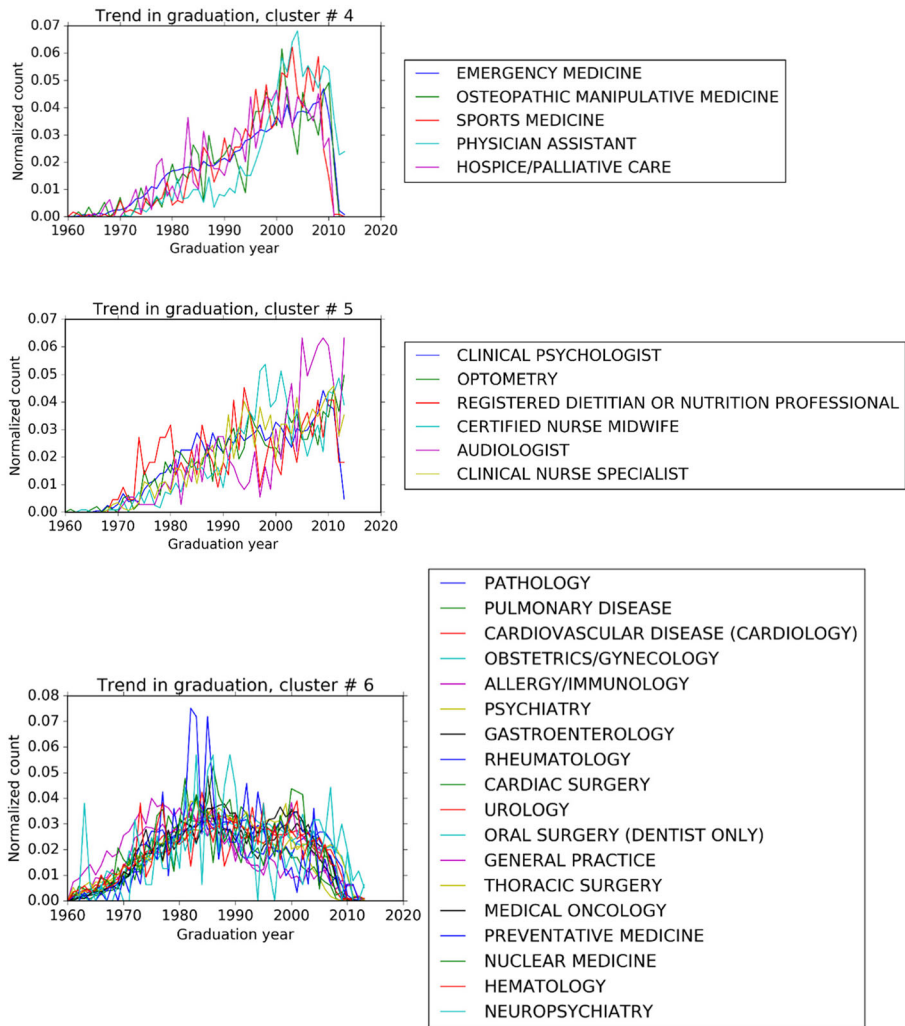
Fig. 6   (continued)

Our analysis in the current paper utilizes data for 895,431 healthcare practitioners, which amounts to 7.2%. Table 1 shows that the CMS data covers more than 75% of current practitioners in the entire USA for several specialties. These percentages are significantly higher than the 3% considered necessary to yield meaningful results [64].

In order to counter the potential criticism in the reader's mind that the hospital total performance score does not reflect individual practitioner performance, the exact same methodology was applied to cardiac surgery outcome data in the New York SPARCS database. The graduation year data was obtained from the New York State Education Department using our custom software. Figure 9 shows that even for specific outcome data, the Spearman's rank correlation coefficient is statistically significant but small.
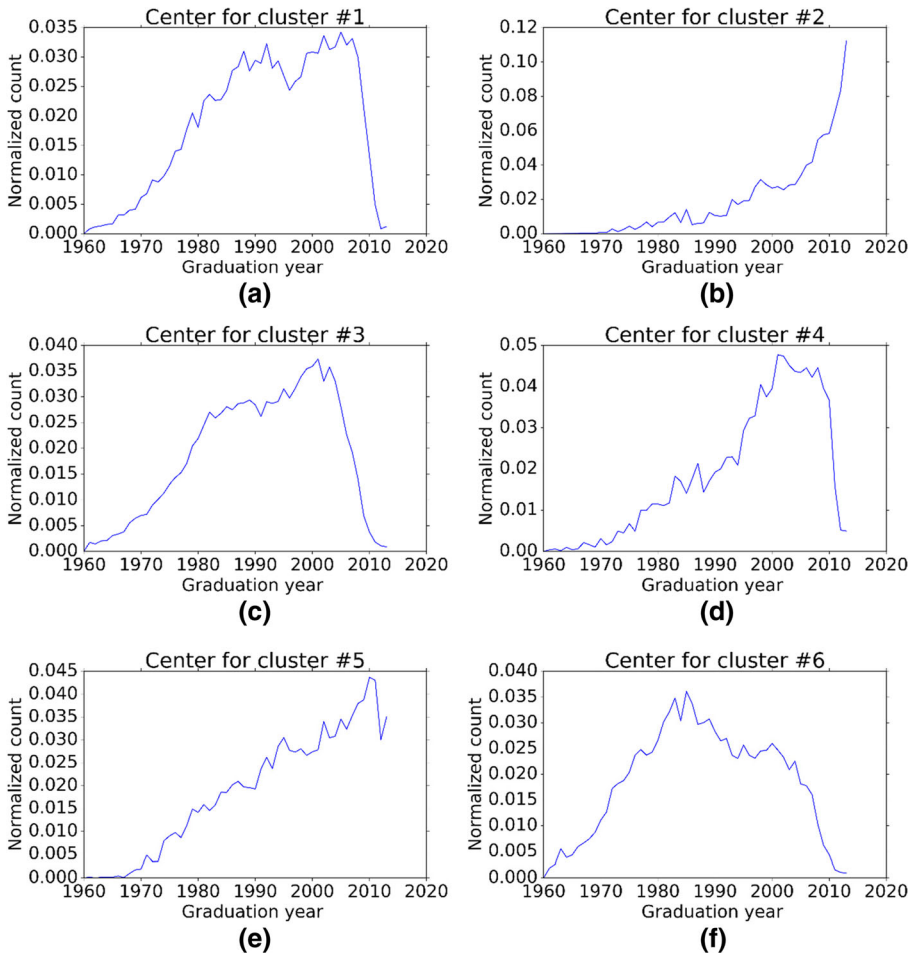
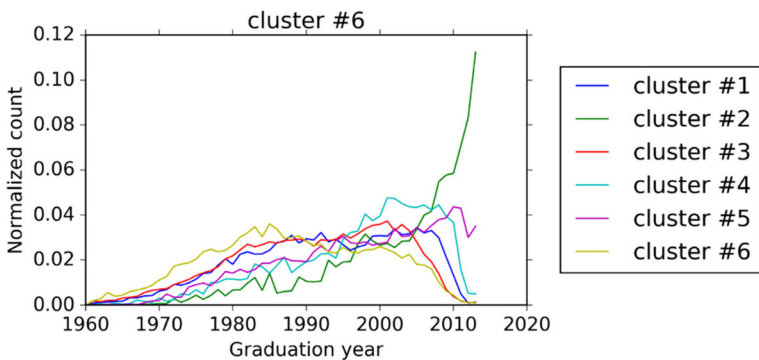**Fig. 7** The cluster centers for each of the clusters shown earlier

Fig. 8 The cluster centers in a single plot to facilitate comparison and may facilitate interpretation of the data by stakeholders. For example, cluster no. 2 compared with all other clusters show growth peak in 2010
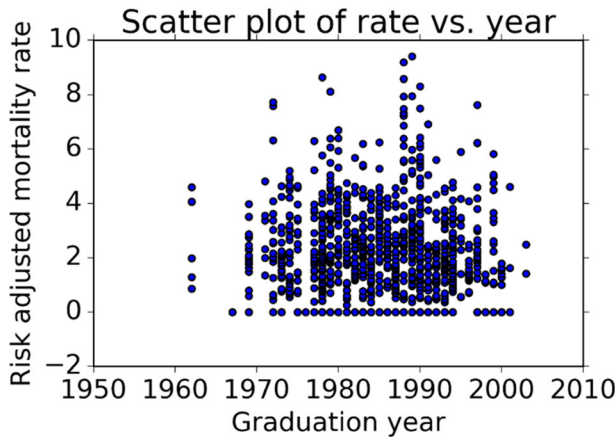
Fig. 9 Scatter plot of risk-adjusted mortality rate versus graduation rate for cardiac surgeons in the NY State SPARCS dataset

This suggests that the seniority of cardiac surgeons determined by graduation year is not likely to affect surgical outcomes.

The framework in this paper illustrates the types of analytics that can be performed and the types of questions that can be considered and answered. The power of this approach consists in utilizing information that is already collected to inform important decisions in healthcare by combining database operations such as JOIN with commands to perform analytics and visualization of the results all within a single-Python-based framework. Further analysis can be performed by applying this freely available framework [11] to more fine-grained data sets.

The current analysis is conducted at a coarse scale of resolution on public data sets with large sample sizes and provides a new perspective on the debate surrounding the factors contributing to outcome measurements. These results suggest that hospitals are well served by the trained healthcare professionals, and changing the seniority mix of these professionals may be unlikely to result in changes in performance metrics. This is only one portion of the puzzle in reining in healthcare costs. Furthermore, it is likely that the rigorous curricula and licensing requirements for healthcare practitioners ensure that anyone who passes the licensing examination is able to provide quality care to patients, at the level expected by the CMS office.

Nurse practitioners are the third largest group in the Medicare database. Figures 3 and 4 show few nurse practitioners graduated around the year 1990, followed by significant growth starting around 2010. It is possible that this trend coincides with the advent of the 2010 Affordable Care Act. This needs to be verified by independent analysis of other data sources. This observation highlights the richness of the CMS data and demonstrates the value of applying the right analytics and visualization techniques.

Figure 4 shows a significant increase in the number of nurse practitioners and nurse anesthetists as compared to family practice and internal medicine, from 2006 to 2014. One reason is that there are more enrollees in Nurse Practitioner programs. A second reason is that nurse practitioners can be deployed immediately after graduation whereas those in Family Practice or Internal Medicine may take more time to be absorbed into full practice.

Figures 6, 7, and 8 demonstrate increasing trends for clusters 2 and 5. Cluster 2 shows an exponential increase and consists of certified registered nurse anesthetist, nurse practitioner, anesthesiology assistant, and physical therapist. Cluster 5 shows a linear increase and consists of specialties including clinical psychologist, audiologist, and clinical nurse specialist. Cluster 1 shows two peaks of graduation, around 1990 and 2005. Clusters 3, 4, and 6 show unimodal distributions, but with different rates of decline. Clusters 3 and 4 show declines after 2000 at different rates. Cluster 6 shows a gradual decline starting in 1980, and includes disciplines such as psychiatry and general practice. These findings have potential policy implications in determining requirements for future generations of practitioners.

Meagher [37] determined that general surgery is one of the specialties that will have the largest projected shortfalls by 2020. Based on our analysis, we can see from Fig. 6 that cluster 3 contains several specialties that are related to surgery, such as general surgery, orthopedic surgery, neurosurgery, vascular surgery, and hand surgery. All of these specialties share the same characteristics, in that the relative number of graduates has been declining steadily since 2000. Furthermore, there are other specialties that share a similar decline, such as diagnostic radiology, geriatric medicine, and pediatric medicine. The value of this result is that it can be readily reproduced from the available data in the CMS repository (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4] by using our open-source analytics pipeline. Furthermore, the analysis can be easily updated as and when new data becomes available every quarter. It would be difficult to reproduce the work conducted by Meagher et al. [37] every quarter, perform it for dozens of specialties and combine all the results in a single repository.

The healthcare landscape is rapidly changing worldwide, with different disease patterns, better treatment options, systematic labor force trends, and evolving cost structures. Accordingly, the curriculum for healthcare providers and public health professionals needs to keep up with these changes, as noted by Fried et al. [67]. The results in our current paper show that it is possible to determine systematic patterns in the healthcare labor force trends through available CMS data (https://data.medicare.gov/Physician-Compare/National-Downloadable-File/s63f-csi6) [4]. Our analysis provides a "bids-eye" view of the trends in healthcare specialties in the USA through a single and comprehensive data source. Furthermore, groups of similarly trending specialties such as nurse practitioner and physical therapy can be utilized for curricular integration, whereby educational institutions could synergize course offerings [38].

Understanding performance outcome data allows for workforce planning to meet future demand for primary healthcare services. For example, 72 million Americans will be 65 and older by 2030, a significant shift in the current demographics of the USA with an increase in chronic disease and morbidity [68]. The affordable care act (ACA) expanded healthcare to 30–34 million uninsured Americans, many through federal Medicaid programs. Selected estimates based on data from the National Health Interview Survey show that the percentage of persons uninsured decreased from 16.0% in 2010 to 9.0% in 2016 (95% confidence interval = 8.43–9.50%) https://www.cdc.gov/nchs/data/nhis/earlyrelease/Earlyrelease201705_01.pdf.

The public release of healthcare data including cardiac surgery outcomes and hospital ratings may affect the selection of providers. For example, Mukamel [69] reported that New York Medicare enrollees were less likely to select a surgeon with

higher published mortality rates. Similarly, Jha and Epstein [70] reported that the odds of ceasing practice was statistically higher for surgeons performing in the bottom quartile.

The framework, approach, and methods in this paper should close the gap between the availability of public healthcare data and the ability to extract meaning and value therefrom. Since the presented tools and techniques have open-sourced by the authors [11], this should spur further analysis, interpretation, and understanding, ultimately lowering the barriers to conduct such research. The use of iPython Notebooks [54] makes it easy for other researchers to understand and replicate the presented results. Python Pandas [53] was released as an open source project in the finance world, and spurred significant research and innovation therein. The world of healthcare could benefit by emulating such a model, with the release of open source projects. The introduction of Electronic Medical Records created challenges such as the lack of inter-operability between different vendor systems [71]. The wave of the future consists of greater transparency across all aspects of healthcare, including open health records [72]. This requires a culture change as noted by Friend [8]: "The 'medical-industrial complex' is not incentivized to share among each other, let alone with patients." The combination of open data sharing as well as code sharing exemplified in the current paper should further accelerate the pace of research and investigation in the healthcare arena. The current national outrage in the USA over escalating drug prices [73] only demonstrates how far we still need to go, as patient advocacy groups appear to be strait-jacketed by accepting significant funding from the same "medical-industrial complex" [42] that they are supposed to defend patients from. In this context, the current paper represents a step in the right direction as open access to data and code [11] will lower barriers to utilizing important healthcare data. A wide audience is reachable by harnessing the well-documented advantages of open source [74] and utilizing attractive visualization front-ends [75]. This should help in addressing one of the shortcomings regarding the public reporting of provider performance, which is the lack of sufficient engagement by consumers [76].

In the past, only hospitals would benefit from the public release of performance data [77]. Based on the framework presented in our paper, several groups can benefit from this information, including educational institutions, policymakers, students interested in choosing appropriate careers and specialties in healthcare, patients seeking to inform themselves about healthcare choices, and citizens concerned about the quality of healthcare.

3. Conclusion

The value of the research presented in this paper lies in the creation of an open source-toolkit that can be freely used by researchers to explore questions in healthcare by using voluminous sources of existing data. This toolkit allows researchers to gain a deeper understanding of trends and relationships among variables that could potentially affect outcomes in healthcare. The toolkit uses Python, Pandas, SciKit and Matplotlib for analyzing, visualizing, and interpreting healthcare datasets.

The relationship between seniority of practitioners and the quality of care they provide is controversial in the healthcare literature. The toolkit was used to analyze data released by CMS and New York State SPARCS. In the CMS dataset, the relationship between graduation year of practitioners and total hospital performance score was weak, based on the Spearman's rank correlation coefficient. Similarly, the relationship between the seniority

of cardiac surgeons as measured by their graduation year and risk adjusted mortality rate of their patients was weak, based on the Spearman's rank correlation coefficient. These results are significant because of the large CMS dataset, representing 7% of all US healthcare workers. These results are different from the norm in other sectors of the economy such as high-technology, where the workers at highly successful companies tend to be younger than workers at other companies.

Color heatmaps provided a detailed visual representation of the distribution of practitioners along the dimensions of graduation year and hospital performance score. This generates a more nuanced view of the correlation scores, and highlights the differences in population trends over a database of 1.2 million practitioners. An unsupervised clustering of graduation trends among specialties identified characteristics of seven major groups, including specialties such as psychiatry that have a slow decrease of practitioners and specialties such as Nurse Practitioners that have an exponential increase in the number of practitioners.

These techniques produce promising results and can be used interactively to explore, analyze, interpret, and visualize large datasets. The application of this technique to data from multiple sources sheds light on factors and trends that can impact education and healthcare policy. The free availability of these tools and techniques should prove valuable to other researchers and software developers in the healthcare arena, and reach a broader audience. This could serve as a new model for conducting healthcare research.

Compliance with Ethical Standards

Conflict of Interest   The authors declare that they have no conflicts of interest.

References

1. Schwartz AL, Landon BE, Elshaug AG, Chernew ME, McWilliams JM (2014) Measuring low-value care in Medicare. JAMA Intern Med 174:1067–1076
2. Schneeweiss S (2014) Learning from big healthcare data. N Engl J Med 370:2161–2163
3. Meyer J, Boll S (2014) Digital health devices for everyone. IEEE Pervasive Comput 13(2):10
4. Medicare.gov The Total Performance Score information. Available: https://www.medicare.gov/hospitalcompare/data/total-performance-scores.html
5. Le ME, Kurd MF (2014) A review and analysis of the YODA trials: what can we glean clinically? Curr Rev Musculoskelet Med 7:189–192
6. Rao AR, Chhabra A, Das R, Ruhil V (2015) "A framework for analyzing publicly available healthcare data," in 2015 17th International Conference on E-health Networking, Application & Services (IEEE HealthCom), pp. 653–656
7. Frakt A (2016) How common procedures became 20 percent cheaper for many Californians, The New York Times
8. Paul R, Clay A (2011) An open source approach to medical research," Stanf Soc Innov Rev
9. Johnson AE, Stone DJ, Celi LA, Pollard TJ (2017) The MIMIC Code Repository: enabling reproducibility in critical care research. J Am Med Inform Assoc p. ocx084
10. Baker M (2016) 1,500 scientists lift the lid on reproducibility. Nature 533:452–454
11. "github.com/fdudatamining/framework, Base Frame Work for Data Mining."

12. Centre For Disease Control and Prevention. Available: http://www.cdc.gov/DataStatistics/
13. New York State Department Of Health, Statewide Planning and Research Cooperative System (SPARCS). Available: https://www.health.ny.gov/statistics/sparcs/
14. Available: http://www.op.nysed.gov/opsearches.htm
15. "Department of Health and Human Services, Secondary Analyses of Existing Data Sets and Stored Biospecimens to Address Clinical Aging Research Questions (R01)."
16. "Department of Health and Human Services, Secondary Dataset Analyses in Heart, Lung, and Blood Diseases and Sleep Disorders (R21)."
17. (2017, 1/28/2018). Robert Woods Johnson Foundation, 2017 Call for Proposals on "Health Data for Action". Available: https://www.rwjf.org/en/library/funding-opportunities/2017/health-data-for-action.html
18. "RUSSELL SAGE FOUNDATION, Funding Opportunity: The Social, Economic, and Political Effects of the Affordable Care Act."
19. Fung CH, Lim Y-W, Mattke S, Damberg C, Shekelle PG (2008) Systematic review: the evidence that publishing patient care performance data improves quality of care. The impact of publishing performance data on quality of care. Ann Intern Med 148:111–123
20. Tsugawa Y, Jha AK, Newhouse JP, Zaslavsky AM, Jena AB (2017a) Variation in physician spending and association with patient outcomes. JAMA Intern Med 177:675–682
21. Higgins A, Brainard N, Veselovskiy G (2016) Characterizing health plan price estimator tools: findings from a national survey. Am J Manag Care 22:126–131
22. Buerhaus P, Staiger D, Auerbach D (2009) The future of the nursing workforce in the United States: Data, trends and implications: Jones & Bartlett Publishers
23. Choudhry NK, Fletcher RH, Soumerai SB (2005) Systematic review: the relationship between clinical experience and quality of healthcare. Ann Intern Med 142:260–273
24. Neumayer LA, Gawande AA, Wang J, Giobbie-Harder A, Itani KM, Fitzgibbons Jr RJ et al (2005) Proficiency of surgeons in inguinal hernia repair: effect of experience and age. Ann Surg 242:344–352
25. Jena AB, Romley J (2015) Changes in hospitalizations, treatment patterns, and outcomes during major cardiovascular meetings—reply. JAMA Intern Med 175:1420–1421
26. Schenarts PJ, Cemaj S (Feb 2016) The aging surgeon: implications for the workforce, the surgeon, and the patient. Surg Clin North Am 96:129–138
27. Anderson BR, Wallace AS, Hill KD, Gulack BC, Matsouaka R, Jacobs JP, Bacha EA, Glied SA, Jacobs ML (2017) Association of surgeon age and experience with congenital heart surgery outcomes. Circ Cardiovasc Qual Outcomes 10:e003533
28. McAlister FA, Youngson E, Bakal JA, Holroyd-Leduc J, Kassam N (2015) Physician experience and outcomes among patients admitted to general internal medicine teaching wards. CMAJ 187:1041–1048
29. Li C-J, Syue Y-J, Kung C-T, Hung S-C, Lee C-H, Wu K-H (2016) Seniority of emergency physician, patient disposition and outcome following disposition. Am J Med Sci 351:582–588
30. Cunningham E, Debar S, Bell B (2003) Association between surgeon seniority and outcome in intracranial aneurysm surgery. Br J Neurosurg 17:124–129
31. Waljee JF, Greenfield LJ, Dimick JB, Birkmeyer JD (2006) Surgeon age and operative mortality in the United States. Ann Surg 244:353
32. Tsugawa Y, Newhouse JP, Zaslavsky AM, Blumenthal DM, Jena AB (2017b) Physician age and outcomes in elderly patients in hospital in the US: observational study. BMJ 357:j1797
33. Melville NA (2017) "Age-based testing of physician competence stirs controversy," in Medscape, ed
34. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L (2006) Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA 296:1094–1102
35. Kupfer JM (2016) The graying of US physicians: implications for quality and the future supply of physicians. JAMA 315:341–342
36. Salthouse TA (2009) When does age-related cognitive decline begin? Neurobiol Aging 30:507–514
37. Meagher AD, Beadles CA, Sheldon GF, Charles AG (2016) Opportunities to create new general surgery residency programs to alleviate the shortage of general surgeons. Acad Med 91:833–838
38. Harden RM (2000) The integration ladder: a tool for curriculum planning and evaluation. Med Educ-Oxf 34:551–557
39. Rao AR, Clarke D (2018) Facilitating the exploration of open health-care data through BOAT: a big data open source analytics tool, in Emerging Challenges in Business, Optimization, Technology, and Industry, ed: Springer, pp. 93–115
40. Bullinger AC, Rass M, Adamczyk S, Moeslein KM, Sohn S (2012) Open innovation in healthcare: analysis of an open health platform. Health Policy 105:165–175

41. Estrin D, Sim I (2010) Healthcare delivery. Open mHealth architecture: an engine for healthcare innovation. Science 330:759–760
42. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Detmer DE (2007) Toward a national framework for the secondary use of health data: an American medical informatics association white paper. J Am Med Inform Assoc 14:1–9
43. Rosenkrantz AB, Doshi AM (2016) Public transparency web sites for radiology practices: prevalence of price, clinical quality, and service quality information. Clin Imaging 40:531–534
44. Krumholz HM (2014) Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 33:1163–1170
45. Rao AR, Clarke D (2016) A fully integrated open-source toolkit for mining healthcare big-data: architecture and applications," in IEEE International Conference on Healthcare Informatics ICHI, Chicago, pp. 255–261
46. Institute of Medicine (2011) Committee on the Robert Wood Johnson Foundation Initiative on the Future of Nursing, The future of nursing: Leading change, advancing health. National Academies Press, Washington, DC
47. Mundinger MO (2002) Twenty-first-century primary care: new partnerships between nurses and doctors. Acad Med 77:776–780
48. Mundinger MO, Kane RL, Lenz ER, Totten AM, Tsai W-Y, Cleary PD, Friedewald WT, Siu AL, Shelanski ML (2000) Primary care outcomes in patients treated by nurse practitioners or physicians: a randomized trial. JAMA 283:59–68
49. New Jersey Division Of Consumer Affairs. Available: http://www.njconsumeraffairs. gov/nur/Pages/APN-Certification.aspx
50. "Summary of the Affordable Care Act," THE HENRY J. KAISER FAMILY FOUNDATION, 2013
51. HHS.GOV/HealthCare. *Read The Law*. Available: https://www.hhs.gov/healthcare/about-the-law/read-the-law/
52. Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95
53. McKinney W (2012) Python for data analysis: Data wrangling with Pandas, NumPy, and IPython: " O'Reilly Media, Inc."
54. Pérez F, Granger BE (2007) IPython: a system for interactive scientific computing. Comput Sci Eng 9: 21–29
55. McKinney W (2010) "Data structures for statistical computing in python," in Proceedings of the 9th Python in Science Conference, pp. 51–56
56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
57. CMS.GOV "Private correspondence with the data vendor for CMS."
58. Pronovost PJ, Lilford R (2011) A road map for improving the performance of performance measures. Health Aff 30:569–573
59. Press, I (2004) The measure of quality. Q Manage Health Care Vol 13, No 4. Lippincott Williams & Wilkins, Inc, Philadelphia, pp. 202–209
60. Kernisan LP, Lee SJ, Boscardin WJ, Landefeld CS, Dudley RA (2009) Association between hospital-reported leapfrog safe practices scores and inpatient mortality. JAMA 301:1341–1348
61. Leonardi MJ, McGory ML, Ko CY (2007) Publicly available hospital comparison web sites: determination of useful, valid, and appropriate information for comparing surgical quality. Arch Surg 142:863–869
62. Rao AR, Clarke D (2017) "An open-source framework for the interactive exploration of big data: applications in understanding healthcare " presented at the IJCNN, International Joint Conference on Neural Networks
63. Physician Specialty Data Book, 2014
64. Hardy Q (2013) Technology workers are young (really young)," The New York Times
65. Thibodeau P (2015) "Median age at Google is 29, says age discrimination lawsuit," Computerworld
66. "Total Healthcare Employment," THE HENRY J. KAISER FAMILY FOUNDATION, 2015
67. Fried LP, Begg MD, Bayer R, Galea S (2014) MPH education for the 21st century: motivation, rationale, and key principles for the new Columbia public health curriculum. Am J Public Health 104:23–30
68. Anderson A (2014) "The impact of the affordable care act on the healthcare workforce," Background
69. Mukamel DB, Weimer DL, Zwanziger J, Gorthy S-FH, Mushlin AI (2004) Quality report cards, selection of cardiac surgeons, and racial disparities: a study of the publication of the New York state cardiac surgery reports. INQUIRY: J Healthc Organ, Provision Financ 41:435–446
70. Jha AK, Epstein AM (2006) The predictive accuracy of the New York State coronary artery bypass surgery report-card system. Health Aff 25:844–855

71. Miller RH, Sim I (2004) Physicians' use of electronic medical records: barriers and solutions. Health Aff 23:116–126
72. Walker J, Darer JD, Elmore JG, Delbanco T (2014) The road toward fully transparent medical records. N Engl J Med 370:6–8
73. Thomas K (2016) "Furor over drug prices puts patient advocacy groups in bind," The New York Times
74. Martínez-Torres MR, Diaz-Fernandez M d C (2014) Current issues and research trends on open-source software communities. Tech Anal Strat Manag 26:55–68
75. DATAUSA. Available: https://datausa.io/about/
76. Hussey PS, Luft HS, McNamara P (2014) Public reporting of provider performance at a crossroads in the United States: summary of current barriers and recommendations on how to move forward. Med Care Res Rev 71:5S–16S
77. Marshall MN, Shekelle PG, Leatherman S, Brook RH (2000) The public release of performance data: what do we expect to gain? A review of the evidence. JAMA 283:1866–1874