EuroVis Workshop on Visual Analytics (2015)
E. Bertini and J. C. Roberts (Editors)

The Use of High-Dimensional Visualizations in Explaining Hospital Pricing Patterns

M. Perkins and G. Grinstein

Institute for Visualization and Perception Research, University of Massachusetts at Lowell, Lowell, Massachusetts, USA

Abstract

The Centers for Medicare and Medicaid Services (CMS) has made public a data set showing what hospitals charged and what Medicare paid for the one hundred most common inpatient stays. By law payments approximate a hospital's cost of providing a service. However, the data shows a wide dollar gap between Medicare payments and what hospitals actually charged. We explore the origins of hospital pricing using the technique of Independent Component Analysis, a form of blind source separation. Discovered source signals were interpreted by putting them into context with conditions, including characteristics of individual hospitals and the marketplaces in which they operate. This high-dimensional data consisting of over 100 variables was explored using Weave, a web-based analysis and visualization environment. Four underlying processes that exert influence on hospital pricing were identified, including one that revealed distinguishing features of hospitals at the extreme high and low ends of the charge distribution. Perhaps surprisingly, hospitals that lie on opposite ends of the price scale were found to have many attributes in common as well.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—General

1. Introduction

Historically, the medical marketplace has lacked price competition - consumers do not have access to information that would allow them to compare what hospitals charge for a given procedure. In May 2014 the Centers for Medicare and Medicaid Services (CMS) publicly released a data set showing what hospitals charged and what Medicare paid for the one hundred most common inpatient stays, for FY 2012 [fMMS13]. The data showed large variations in hospital charges for similar services, with regional patterns but also a wide price range within the same geographic area.

Hospital charges are the starting point for negotiated payments, which ultimately depend on whether the patient is insured and the insurer's ability to contract for discounted prices with the hospital. Patients who are uninsured or who have exceeded their insurance policy's coverage limits may be billed undiscounted charges [Bri13]. The practice of cross subsidization occurs widely within the health care system, meaning that profit maximizing prices are charged to some payers or for some services and the surpluses are used to subsidize other payers or other clinical services [Com04]. This practice may encourage hospitals to use pricing that is decoupled from procedure-related costs of care.

In this paper we describe our effort to identify the factors that exert greatest influence on hospital pricing through the novel application of Independent Component Analysis (ICA), a mixed signal decomposition technique, to the CMS hospital charge data. Particular discoveries about the underlying source processes came out of the combined use of ICA and visualization. This study was done using Weave [BDK*11, BSD*09, BSR*11, DFSG12] and Weave Analyst [CSG*13, IVR15a], a new, comprehensive data analysis and presentation system developed at the University of Massachusetts Lowell's Institute for Visualization and Perception Research (IVPR) and freely available on Github [IVR15b].

2. Characteristics of the CMS Provider Inpatient Data – Preliminary Explorations

The data set used in this study includes FY 2012 hospital-specific charges for 3,317 U.S. hospitals and corresponding payments made by Medicare for the 100 most frequently

DOI: 10.2312/eurova.20151099

billed discharges. Charges are determined by hospitals for items and services provided to patients. When a hospital contracts with Medicare, it agrees to receive a predetermined fixed amount as payment in full.

The overall correlation coefficient between hospital charges and Medicare payments as calculated from the CMS data is 76.5%, indicating a strong though not complete relationship between what hospitals actually charge and their costs. However the dollar gap is great – total hospital charges were 4.7 times that of Medicare payments. These relationships are illustrated in Figure 1, for a sample of four Diagnostic Related Groups (DRGs) with different inpatient resource requirements. Individual hospitals are represented as points in the scatterplot and colored by their charge-to-payment ratio. The proportion of hospitals in each bin is similar across DRGs, with the greatest concentration of hospitals charging 2-4 times Medicare payments, but with many charging far higher. The correlation coefficients range between 23.7%-35.8%, which is weaker at the procedure level than overall and indicative of cross-subsidization. Such information emphasizes the need to further explore the data and identify factors that lie at the source of hospital pricing.

A first, straightforward observation is that DRG charge levels exhibit distinct geographic patterns. The variation by Hospital Referral Region (HRR) is displayed in Figure 2 for "DRG 178 – Respiratory Infections & Inflammations W CC". HRRs represent regional health care markets for tertiary medical care that generally requires the services of a major referral center. Price influences at the HRR level may include broad market and demographic factors. Hospital pricing also varies within region, as shown for the New York City area in Figure 3. Average charges for cases assigned to the same DRG in the mapped area range from $6,371 to $172,875. Price influences at this level may include hospital specific attributes and micro-market characteristics.

3. Independent Component Analysis

ICA generates a model for the observed multivariate data. In this model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed non-Gaussian and mutually independent, and are called the independent components of the observed data. ICA has been applied in a diverse range of fields, to analyze digital images, economic indicators, financial time series, and biomedical signals [NK11].

Our goal is to find the "original source signals" influencing hospital pricing. We are limited by being able to observe only mixtures of the signals, as represented by the vectors of hospital average charges by DRG. The vectors constitute data points in 100-DRG dimension space. The ICA method centers the data about the origin and then removes any correlations in the data through a linear change in the coordinates. This restores the original "shape" of the data before mixing of the signals. ICA then separates sources by rotating the n-dimensional axis such that the non-Gaussianity of the data projection on each axis is maximized [HO00]. The ICA framework can be represented as matrix product X = SA where X is the observed data, S contains the independent components and A is the transform matrix from the source space S to the data space X.

ICA was performed using R's "fastICA" package. To find the independent components, non-Gaussianity is measured using approximations to negative entropy. The algorithms for component extraction are detailed in [Hyv99].

Figure 1: *Full size image with caption*

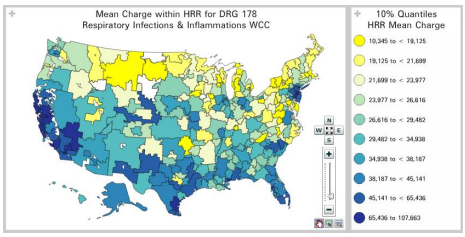

Figure 2: *Full size image with caption*

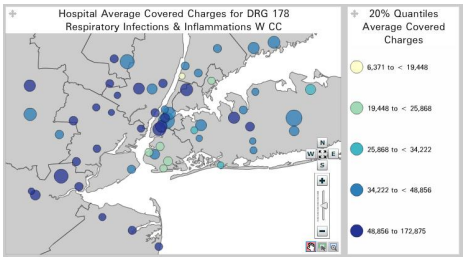

Figure 3: *Full size image with caption*

3.1. Application of ICA to the Hospital Charge Data

The CMS data contains the one hundred most frequently billed DGRs, with a minimum of 11 discharges required

for reporting. As billed DRGs often did not achieve 11 discharges, the 3317 x 100 matrix representing the original data contained missing values. These were imputed using the DRG mean charge across all hospitals to generate X. Since ICA centers the data by subtracting the column means, we were able to utilize all of the information in the data set without introducing bias.

We define the hospital DRG charge index as the ratio of the hospital's Average Covered Charges for that DRG to the mean charge across all hospitals. Restating the charge level as a relative rather than absolute dollar amount standardizes the scale of the variables and allows for extracted sources to be interpreted independent of procedure type. We define the hospital mean charge index, the hospital's overall charge level, as the mean of its charge indices across DRGs. Note that imputed values were included in the calculation of this metric. An alternate treatment could have been to exclude non-reported DRGs for the hospital, which would have produced a wider range of hospital mean charge indices, and a greater tendency for small volume hospitals to appear as outliers.

4. Results

From the application of ICA to the hospital charge data we discovered source signals underlying the hospital charge profiles. In order to interpret their meaning, we selected the hospitals whose value in the considered source exceeds +/- 2.5% of the distribution, thereby forming coherent groups. We next put these groups into correspondence with already-known characteristics of the hospitals and marketplaces in which they operate. Finally, we used Weave to visualize the relationship between the hospital groups, their charge levels, and distinguishing indicator variables.

4.1. Weave Analyst

Weave Analyst is an integration of the Weave data visualization platform with R. It enhances the data warehousing and analytical capabilities of the Weave environment, allowing users to analyze, visualize and report data in an integrated framework. For this study, the hospital charge data along with additional public source data sets were imported into Weave making them accessible through Weave Analyst. The project's R scripts, which varied in complexity based on the needs of the exploratory and analytical tasks, were selected through the user interface and run against the integrated data tables. Results were automatically returned to Weave where they were explored using interactive linked visualizations. A snapshot from Weave Analyst is show in Figure 4.

4.2. Visualization of the Independent Component Group Member Hospitals

Selection of the hospitals in the tails of the independent component distribution resulted in two groups of hospitals (ex-

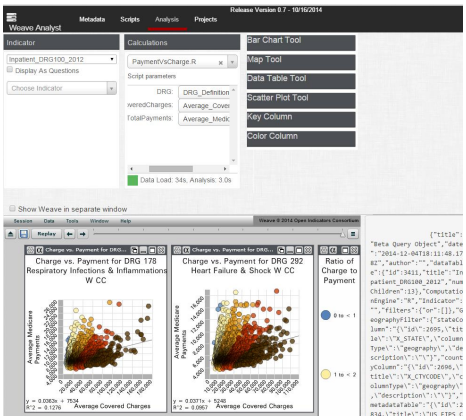

Figure 4: *Full size image with caption*

pressed and anti-expressed) for each of the four independent components. As can be seen in Figures 5a and 5b, the hospital groups identified by ICA typify different pricing levels. There is overlap between some groups as a hospital may be strongly influenced by more than one source signal. This is illustrated in Figure 5a. The overall mean charge index of each section of the Venn diagram (IC groups and their intersections) is represented by a relative color level, and the numerical values displayed in the color legend.

Individual hospitals were plotted vs. their mean charge index, as shown in Figure 5b. Bar height corresponds to hospital overall discharge volume. When an area of the Venn diagram is selected, individual hospitals are highlighted on the index. The source signals, which are ambiguous as to scaling and sign, are thus put into context according to where member hospitals cluster on the charge scale.

Displaying the hospitals on a map puts the IC groups into a more intuitive context with demographic variables, as shown in Figure 5c. A countrywide view enables all of the hospital points to be seen at once, and also for the geographic concentration of selected IC groups to be compared.

4.3. Interpretation of Source Signals

Putting the IC group member hospitals into context with conditions is a technique by which the source signals may be interpreted in terms of processes which govern pricing. ICA has been similarly used towards the identification of characteristic gene expression patterns related to distinct biological functions [Lie02, CRT04]. Hospital information for FY 2012 was obtained from the CMS cost report [fMMS12a] and sub-reports on wage index and case mix data [fMMS12c, fMMS12b], along with quality of care outcomes from the Hospital Compare database [fMMS12d]. Demographic information was obtained from the U.S. Census Bureau American Community Survey [Bur12]. In total,

136 data fields were collected for each of the significantly expressed hospitals.

Many of the fields were able to be predicted based on other variables, a phenomenon known as multicollinearity. In order to facilitate source signal interpretation we reduced the dimensionality of the data using a correlation-based filter. For the resulting set of predominant features, we examined the t-statistic comparing the IC group mean to all other IC groups combined. The final set after these refinements contained 51 "indicator" variables.

A heat map visualization is used to discern the indicator makeup of the hospitals in each IC group, as shown in Figure 5d. IC groups are ordered along the x-axis by increasing mean charge index. Indicator variables are shown on the y-axis. The heat map is colored according to the value of the t-statistic. As can be seen, the hospital groups identified by ICA typify different levels of the indicator variables. The heat map gives an overall impression that hospital charges increase with increasing levels of most of the indicator variables. Purple color in the heat map signifies a negative t-statistic, i.e. the mean of an indicator variable is lower for that IC group than for all other IC groups combined, while orange signifies a higher mean for the IC group.

We employed Weave's capabilities to interpret the source signals as follows:

- The dual areas of an IC group were highlighted in the Venn diagram, for example IC1.Grp1 and IC1.Grp2. Hospitals in these two groups are expressed and anti-expressed in the IC1 source signal.
- Member hospitals of those IC groups were automatically highlighted in the charge scale and geographic map. The characteristic t-statistics of the IC groups were likewise highlighted in the heat map.
- The charge levels of IC group member hospitals and their geographic locations were noted.
- Contrasting indicator variables between the two t-statistic columns in the heat map were identified.

IC2 proved to be the most interesting of the four discovered source signals. IC2.Grp1 has the lowest mean charge index of any of the hospital groups at 0.524, while IC2.Grp2 has the highest charge at 1.829. The low charge group is tightly clustered on the charge scale, and the hospitals are located in urban areas with concentrations in the Northeast and Midwest. The high charge group is located in major metropolitan areas. Both groups are large volume hospitals.

Figure 6 compares the expressed and anti-expressed hospitals in IC2. The two groups have many attributes in common – they are large teaching hospitals, performing a relatively high proportion of surgical DRG procedures, and have mostly good performance scores (the high charge group has better scores overall). The low charge group has smaller networks, a low average length of stay, and low occupation adjusted hourly wages. The high charge group has an extremely low ratio of outpatient to inpatient revenue, and its hospitals tend to be referral centers performing specialized care. Type of control tends towards governmental and voluntary not-for-profit in the low charge group, vs. proprietary in the high charge group.

Figure 5: *Full size image with caption*

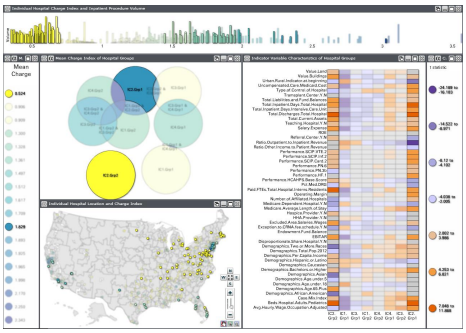

Figure 6: *Full size image with caption*

The detailed results for each of the source signals and source signal interactions are beyond the scope of this paper. Here we have described the use of integrated visualization and analytics as applied to the CMS hospital charge data. This high dimensional data was able to be explored with Weave Analyst, with resulting insights to the underlying processes that govern hospital pricing. The interactive visualizations are available as a Weave demo.

5. Conclusions

Ordering the IC groups by mean charge resulted in a visual progression of the indicator variable levels. To a large extent, hospital pricing tracked with known influences on hospital costs such as those used in the Medicare payment formulas. As far as explaining the wide gap between Medicare payments and hospital charges, a clue is provided by the interpretation of the source signal that underlies pricing at the high and low extremes. An area of future research could be to incorporate additional data sources that expand the perspective on conditions in which these hospitals operate.

References

[BDK*11] BAUMANN A., DUFILIE A., KOLMAN S., KOTA S., GRINSTEIN G., MASS W.: Exploratory to presentation visualization, and everything in-between: Providing flexibility in aesthetics, interactions and visual layering. In *Proc. IV '11* (2011), pp. 200 – 204. doi:10.1109/IV.2011.82. 1

[Bri13] BRILL S.: Bitter pill: why medical bills are killing us. *Time 2* (2013). 1

[BSD*09] BAUMANN A., SMRTIC M., DUFILIE A., MASS W., GRINSTEIN G.: Experiences in the development of a measure and indicator web-based visualization system. *VisWeek* (2009). 1

[BSR*11] BAUMANN A., SHAMS S., ROSS M., MASS W., GRINSTEIN G.: Enhancing stem classes using weave: A collaborative web-based visualization environment. In *Proc. ISEC '11* (2011), pp. 2A–1 – 2A–4. doi:10.1109/ISECon.2011.6229637. 1

[Bur12] BUREAU U. S. C.: American community survey. http://www.census.gov/acs/www/data_documentation/2012_release/, 2012. 4

[Com04] COMMISSION F. T.: Improving health care: A dose of competition: A report by the federal trade commission and the department of justice. http://www.ftc.gov/reports/improving-health-care-dose-competition-report-federal-trade-commission-department-justice, 2004. 1

[CRT04] CHIAPPETTA P., ROUBAUD M., TORRESANI B.: Blind source separation and the analysis of microarray data. *Computational Biology 11*, 6 (2004). 3

[CSG*13] CRAWFORD C. G., SMYSER M., GRINSTEIN G., RIBBLE J., PARK S., CHAPMAN R., PURUSHE S., RYAN P., KAMAYOU F., GALKINA E.: Visualizing health: enhancing public health through weave data analysis and visualization. In *Proc. IEE VIS '13* (2013). 1

[DFSG12] DUFILIE A., FALLON J., STICKNEY P., GRINSTEIN G.: Weave: A web-based architecture supporting asynchronous and real-time collaboration. In *Proc. AVI '12* (2012). 1

[fMMS12a] FOR MEDICARE & MEDICAID SERVICES C.: Cost reports details for title: 2012. http://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Cost-Reports/Cost-Reports-by-Fiscal-Year-Items/HOSPITAL10-DL-2012.html?DLPage=1&DLSort=0&DLSortDir=descending, 2012. 3

[fMMS12b] FOR MEDICARE & MEDICAID SERVICES C.: Fy 2012 final rule data file. http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY-2012-IPPS-Final-Rule-Home-Page-Items/CMS1250507.html, 2012. 4

[fMMS12c] FOR MEDICARE & MEDICAID SERVICES C.: Fy 2012 wage index home page. http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Wage-Index-Files-Items/CMS1239640.html, 2012. 4

[fMMS12d] FOR MEDICARE & MEDICAID SERVICES C.: Official hospital compare data. https://data.medicare.gov/data/hospital-compare, 2012. 4

[fMMS13] FOR MEDICARE & MEDICAID SERVICES C.: Inpatient Charge Data FY 2012 kernel description. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient2012.html, 2013. 1

[HO00] HYVÃĎRINEN A., OJA E.: Independent component analysis: A tutorial. *Neural Networks 13*, 4-5 (2000), 411–430. 2

[Hyv99] HYVÃĎRINEN A.: Fast and robust fixed-point algorithms for independent component analysis. *Transactions on Neural Networks 10*, 3 (1999), 626–634. 2

[IVR15a] IVRP/WEAVE: Weave (web-based analysis and visualization environment). http://iweave.com/index.php?page=aws, 2015. 1

[IVR15b] IVRP/WEAVE: Weave (web-based analysis and visualization environment). https://github.com/IVPR/Weave, 2015. 1

[Lie02] LIEBERMEISTER W.: Linear modes of gene expression determined by independent component analysis. *Bioinformatics 18* (2002), 51–60. 3

[NK11] NAIK G., KUMAR D.: An overview of independent component analysis and its applications. *Informatica 35*, 1 (2011), 63–81. 2