NYS SPARCS Public 2014 Dataset: Let's Look at Septicemia, What Can We Learn?

Lisa Szydziak, MS[1], Simon Sheather, PhD[2]
[1]Northwell Health – Southside Hospital, 301 E. Main St, Bay Shore, NY 11706
[2]University of Kentucky, Lexington, KY 40506

Abstract

New York State Department of Health provides publicly available de-identified data for the general public and researchers at www.health.ny.gov/statistics/gov referred to as SPARCS data. SEPTICEMIA & DISSEMINATED INFECTIONS cases representing a high volume APRDRG in the 2014 SPARCS dataset were investigated. A statistical analysis enables a comparison of facilities in terms of charges and costs adjusting for race, gender, age, length of stay (LOS), admit type, disposition, Severity of Illness (SOI), Risk of Mortality (ROM), payer, procedure, high cost, admit day of week and discharge day of week. In modeling charges/costs, a single model was found not to fit and models differed across 2 groups: "Patients that die in 1 day" and "Patients that do not die in 1 day". A Logistic Model was produced for examining the probability of death in "Patients that do not die in 1 day" adjusting for gender, age, LOS, admit type, disposition, SOI, ROM, facility, payer, procedure and discharge day of week. Modeling of publicly reported data is very powerful, it could support policy makers, administrators and patients in understanding the complexity of healthcare costs and quality.

Key Words: Septicemia, SPARCS, statistical analysis, charges, costs, septicemia mortality

1. Introduction

One of the most common reasons for inpatient hospitalization is Sepsis, a potential life-threatening complication of an infection which can result in death. This study sought the opportunity to analyze the New York State SPARCS (Statewide Planning and Research Cooperative System) database to understand what drives costs, charges and mortality associated with SEPTICEMIA & DISSEMINATED INFECTIONS. Under the NYS Open Data initiative, the government allowed the public access to datasets to analyze and explore in an effort to be innovative, transparent and consumer-focused.[1] New York State offers an Inpatient de-identified dataset (SPARCS data). This file contains basic record level detail for the discharge, i.e. patient characteristics, diagnoses, treatments, services, charges and costs.[2]

Charges, costs and payments relating to a hospital stay can cause confusion to a patient. In fact, these three terms are different and often misunderstood. Charges are the list

[1] NYS-ITS. 2016. *Data Submission Guide*. 2016. https://data.ny.gov. Accessed February 14, 2018, 4.

[2] NYSDOH Bureau of Health Informatics. *SPARCS Hospital Inpatient Discharges De-identified File with Cost*. 2014. https://www.health.data.ny.gov. Accessed February 14, 2018. 2.

prices, maintained on a list called the "chargemaster" which varies by hospital. The government and private insurers rarely pay full charges, and in fact, uninsured and underinsured patients will often receive a charitable discount. Costs are expenses relating to the hospital stay such as room and board, pharmaceuticals, and supplies as well as administrative expenses. Lastly, the payment received by patients, government and insurers is the amount the hospital actually receives. The payment is fixed by the government (Medicare) based on diagnosis and procedures, or negotiated directly with private insurers, or received directly from the patients via cost-sharing and balance billing. Furthermore, there is a portion of care that is uncompensated and referred to as charitable care where the hospital may receive no payments.[3]

We will examine the variation in hospital charges and costs relating to this high volume illness, Septicemia, in New York State. Furthermore, what are the most significant drivers predicting death in those Septicemia patients?

2. Material and Methods

2.1 Setting

The publicly available 2014 Hospital Inpatient discharges (SPARCS de-identified) dataset contains 2,365,208 records for 215 facilities. From this dataset, there were 84,721 APR DRG: "SEPTICEMIA & DISSEMINATED INFECTIONS" discharge records in 191 facilities in New York State included in the analysis. This study does not require IRB approval because it is an analysis of a publicly available de-identified dataset.

2.2 Data Preparation

The following patient characteristics were used as predictor variables in the multiple linear regression equation: facility, age group, gender, race, length of stay (LOS), admit day of week, patient disposition, discharge day of week, procedure, Severity of Illness (SOI), Risk of Mortality (ROM), payer, admission type with response variables of total charges and total costs.

Total charges are defined as billed amounts and vary by facility for a variety of reasons: hospital, payer mix, comparability of billing, physician judgment, outlier cases, quality of care and region. Costs represent the cost of the care and vary by facility as well. Total costs are <u>estimated</u> by a facility specific RCC (ratio of cost to charges) as reported to NYS in the ICR (Institutional Cost report)[4]. The "Total Costs" may provide a consistent way to compare hospitals, however, estimated costs are derived and may not necessarily represent the final cost of the service.[5]

The patient's status upon discharge is the Patient disposition. We collapsed the "patient disposition" variable from 19 dispositions to the following 6 dispositions: another

[3] Kahn, C. *Words Matter: Defining Hospital Charges, Costs and Payments - And the Numbers that Matter Mosts to Co.* 2015. https://fah.org/blog/. Accessed February 14, 2018.

[4] NYS DOH Bureau of Primary and Acute care reimbursement. 2010 Institutional Cost Report. November 2012. https://www.health.data.ny.gov. Accessed February 14, 2018.

[5] NYSDOH Bureau of Health Informatics. *SPARCS Hospital Inpatient Discharges De-identified File with Cost*. 2014. https://www.health.data.ny.gov. Accessed February 14, 2018.

hospital facility, expired, home, hospice, left against medical advice (AMA), SNF. We also collapsed 10 payers to 5 payers: health insurer, Medicaid, Medicare, self-pay, unknown.

2.3 Modeling Charges, Costs and Live/Die from Patient Characteristics

We log-transformed the following variables: charges, costs, length of stay (LOS). A procedure variable TRT_DX was created based on the procedure variable and was mapped to either “Treatment”, “Diagnosis”, or “None”. Furthermore, we have defined “Death” as disposition of “Expired” or “Hospice”, given that “Hospice” expected outcome is death.

“Patients that Die within 1 day” were found to exhibit extraordinarily high or low charges. Perhaps this phenomenon can be explained by the following scenarios. Patients who die within one day are extremely sick and their care can widely vary. Some patients have little medical intervention due to a DNR (do not resuscitate) resulting in low charges /costs. On the other hand, extreme measures were employed to try to save this patients life by way of procedures and treatments, tests which account for high charges/costs.

A high/low categorical variable for “Patients that Die in 1 DAY” was created for both charges and costs: D1HC = high charges (>30000), D1HC = low charges (<=30000); D1HCosts = high costs (>10000), D1HCosts = low costs (<=10000). There is a similar delineation between high/low charges (Figure 1) and high/low costs (Figure 2).

Figure 1: Charges: Low - High

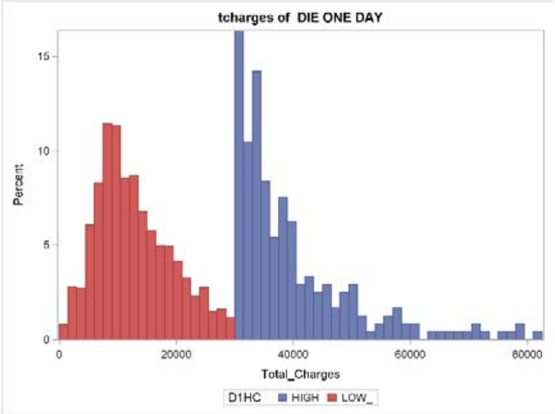
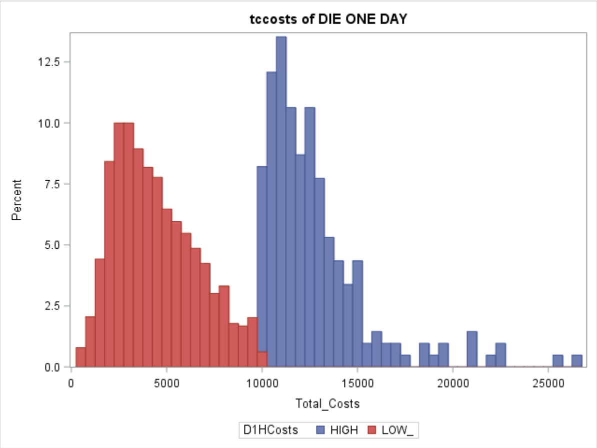


Figure 2: Costs: Low - High



Upon exploration of the data, we found records with 1) costs > charges and 2) high LOS with inordinately low costs/charges. Neither of these situations is sensible, and were excluded from the analysis.

To analyze the data we will use separate multiple linear regression models for log(charges) and log(costs). In modeling charges/costs, a single model was found not to fit and differed across 2 groups: “Patients that die in 1 day” and “Patients that DO NOT die in 1 day”. Separate models were fit for “Patients that die in 1 day” and the remaining population, “Patients that DO NOT die in 1 day” (die on day 2 or later, or survive) for charges/costs.

The dependent variables are log(charges) and log(costs) respectively. Predictor variables were race, gender, age, log(LOS), admit type, patient disposition, Severity of Illness (SOI), Risk of Mortality (ROM), facility, payer, admit day of week, procedure, discharge day of week. The “Patients that die in 1 day” models contain a high/low cost/charge variable.

Once we run the analysis, marginal model plots are utilized to provide information comparing the fitted model to a nonparametric fit. It may be necessary to add a quadratic term for log(LOS) to obtain a better model fit. Fit diagnostics such as residuals, outliers, studentized residuals, outliers, leverage points, quantile plots, and CooksD points are assessed for model validity.

In summary, we will discuss 4 Multiple Linear regression models: 1) Charges for “Patients that die in 1 day”,   2) Costs for “Patients that die in 1 day”, 3) Charges for “Patients that DO NOT die in 1 day”, and 4) Costs for “Patients that DO NOT die in 1 day”.

A reduced model based on the significant predictor variables was then run. A partial F test will compare the full and reduced model for the same response variable and reveal whether or not it is prudent to use the more parsimonious model. Inherent in the model is the ability of the predictor variables to be sorted in order of importance, at which point the key drivers of the response variable will become evident.
In addition, a logistic regression model examining the probability of death adjusting for gender, age, LOS, Type, SOI, ROM, facility, payer, discharge day of week, TRT_DX was derived. This model was developed with a training and validation set.

3. Results

3.1 Model I: Total Charges for "Patients that die in 1 day"

Our first model's response variable is log(charges) for the population N=3,127 of "Patients that die in one day". A partial F test confirms the validity utilizing a reduced model with the following predictor variables and their corresponding p-values (Table 1).

Table 1: Overall Effect Tests: Multiple linear regression model predicting charges - "Patients that die in 1 day"

Source	DF	Type III SS	Mean Square	F Value	Pr > F
D1HC	1	91.12	91.12	688.75	<.0001
TRT_DX	2	46.11	23.05	174.26	<.0001
Type_of_Admission	3	61.75	20.58	155.58	<.0001
SOI	3	8.89	2.96	22.4	<.0001
Age_Group	4	9.01	2.25	17.03	<.0001
Facility_Name	164	281.79	1.72	12.99	<.0001
Payer	4	2.48	0.62	4.69	0.0009
ROM	3	1.70	0.57	4.28	0.0051

The variable which describes high or low charges is highly significant. Patients who die within one day are extremely sick and their care can widely vary. Some patients have little medical intervention (perhaps a DNR) resulting in low costs/charges. On the other hand, a no-holds-barred approach is employed to try to save this patients life which account for high charges/costs. The variable TRT_DX representing procedures will affect the charges. The admission type is a strong factor in predicting charges with an emergency admission being most costly. SOI, Age, facility, payer and ROM are significant factors as well.

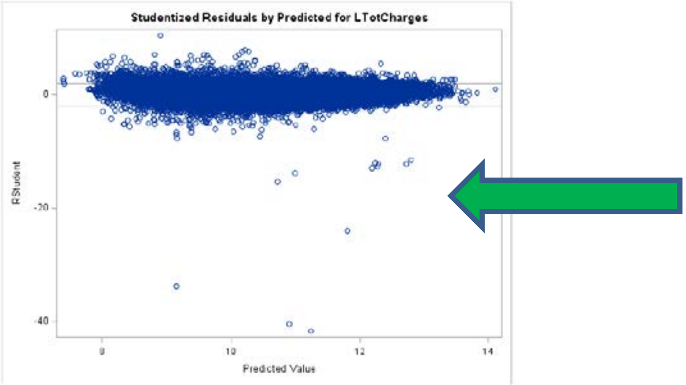
3.2 Model II: Total Costs for "Patients that die in 1 day"

The response variable is log(costs) for the population N=3,127 of "Patients that die in one day". A partial F test confirms the validity utilizing a reduced model with the same predictor variables used in model I. The order of significance remains the same, with the exception of Payer and ROM swapping places. The similar results of Model I log(charges) and Model II log(costs) are not surprising because costs and charges are related.

3.3 Model III: Total Charges for "Patients that DO NOT die within 1 day"

The response variable is log(charges) for the population of “Patients that DO NOT die in one day” (N=81,529). Preliminary analysis revealed nonsensical data points: Short stays with inordinately high charges/costs, or long stays with super low charges/costs. These points were deemed erroneous and removed from model III and IV analyses (Figure 3).

Figure 3: Studentized Residuals.



Initial predictor variables are: race, gender, age, log(LOS), admission type, disposition, SOI, ROM, facility, payer, admit day of week, discharge day of week, TRT_DX. The marginal model plot prompted concern to add a quadratic variable for log(LOS)^2 which improved the model fit. Log(LOS) and log(LOS)^2 were the most important variables driving charges (Table 2).

Table 2: Overall Effect Tests: Multiple linear regression model predicting charges - “Patients that DO NOT die in 1 day”

Source	DF	Type III SS	Mean Square	F Value	Pr > F
log(LOS)	1	1098.08	1098.08	16824.40	<.0001
(log(LOS))^2	1	212.71	212.71	3258.99	<.0001
TRT_DX	2	338.02	169.01	2589.51	<.0001
Facility_Name	186	12817.08	68.91	1055.80	<.0001
ROM	3	172.82	57.61	882.65	<.0001
Disposition	5	159.44	31.89	488.57	<.0001
SOI	3	87.95	29.32	449.20	<.0001
Age_Group	4	90.08	22.52	345.05	<.0001
Type_of_Admission	5	39.14	7.83	119.93	<.0001
Discharge_Day_of_Week	6	5.66	0.94	14.46	<.0001
Gender	1	0.94	0.94	14.34	0.0002
Race	3	1.41	0.47	7.18	<.0001
Admit_Day_of_Week	6	1.46	0.24	3.74	0.001
Payer	4	0.75	0.19	2.87	0.0217

The TRT/DX procedure variable impacts the charges. Furthermore, differences in charges can be attributed to the facility, a facility can charge whatever they want, and this model does capture that sentiment. The ROM, SOI and disposition tell us how sick a patient is and the sicker, the higher the charges. The dramatic difference of the coefficient in the high (1.97) versus low (-0.56) facility is noteworthy (Table 3).

Table 3: Parameter estimates top 5 significant predictors: Multiple linear regression model predicting charges - "Patients that DO NOT die in 1 day"

Parameter	Estimate	Standard Error	t Value	Pr>|t|
log(LOS)	0.58	0.00	129.71	<.0001
$(\log(LOS))^2$	0.06	0.00	57.09	<.0001
TRT_DX DX	0.38	0.03	13.91	<.0001
TRT_DX TRT	0.37	0.02	18.21	<.0001
TRT_DX NO PROC	0.00	-	-	-
Facility (Low)	-0.56	0.07	-7.76	<.0001
Facility (High)	1.97	0.04	48.35	<.0001
ROM 4 Extreme	0.25	0.01	42.65	<.0001
ROM 3 Major	0.08	0.00	19.18	<.0001
ROM 2 Moderate	0.02	0.00	12.64	<.0001
ROM 1 Minor	0.00	-	-	-

3.4 Model IV: Total Costs for "Patients that DO NOT die within 1 day"

The response variable is log(costs) for the population of "Patients that DO NOT die in one day" and yields similar results as Model III with the order of importance of the top 5 variables as follows: log(LOS), TRT_DX, $[\log(LOS)]^2$, ROM, facility.

Cost is primarily driven by how long a patient stays in the hospital. Procedures impact the cost. Facilities have differing costs that can be a function of available resources, services and differing physician practice. Cost is also driven by the risk of mortality and severity of illness, which tells us, the sicker the patient, the higher cost.

We find similar results of Model III (charges) and Model IV (costs) for the population of "Patients that DO NOT die in 1 day". As a matter of fact, the order of variables sorted by significance is alike in both models with a couple of variables swapping places.

3.5 Model V: Logistic Model predicting Death for "Patients that DO NOT die within 1 day"

A logistic model was produced for examining the probability of death in "Patients that DO NOT die in 1 day" adjusting for gender, age, LOS, admission type, disposition, severity of illness, risk of mortality, facility, payer, procedure, discharge day of week. A training and validation set was utilized to derive this model. Death is defined as a patient disposition of "Expired" or "Hospice," given that Hospice expected outcome is death. To overcome the issue of quasi-complete separation, low volume facilities (Expiry<6 or

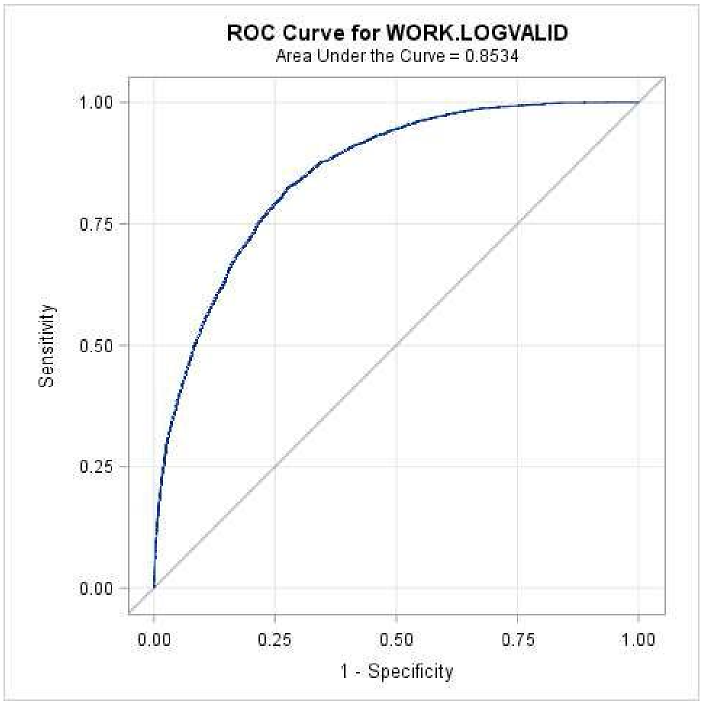
Cases<100) were excluded. The variables sorted in order of significance are: ROM, log(LOS), SOI, Facility, discharges day of week, age, TRT_DX, admission type, payer, gender. A partial list of odds ratios reveals the Risk of Mortality (ROM), log(LOS), and Severity of Illness (SOI) as key variables in predicting mortality (Table 4).

Table 4: Partial list Odds Ratios and 95% CL: Logistic Model predicting death – "Patients that DO NOT die in 1 day"

Odds Ratios			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
ROM 2 Moderate vs 1 Minor	9.29	4.45	19.38
ROM 3 Major vs 1 Minor	40.36	19.36	84.14
ROM 4 Extreme vs 1 Minor	120.88	57.81	252.77
Log(LOS)	0.52	0.50	0.54
SOI 4 Extreme vs 1 Minor	4.08	1.99	8.34

The ROC curves exhibited c=.8532 for the training set and c=.8534 (Figure 4) for the validation set.

Figure 4: ROC Curve for Validation dataset.



3.6 Limitations

The model was restricted to variables available in the SPARCS de-identified data set. There may be variables in the identifiable data set that may better predict charges/costs and mortality.

4. Discussion

Sepsis, a serious complication of septicemia, is a common diagnosis associated with high costs and often fatal outcomes. The elderly are particularly susceptible to this disease, defined as a systemic inflammatory syndrome in response to infection associated with acute organ dysfunction. There have been many therapies, protocols and research efforts devoted to understand and treat this very serious and commonplace disease.[6] The statistical models explore the significant drivers of hospital charges, costs and mortality of SEPTICEMIA & DISSEMINATED INFECTIONS in New York State.

However, before we review our results, it is important to expand upon the often misunderstood differences of price by facility. Charges vary predominately across hospitals which are driven by differences of hospital characteristics. The “chargemaster” is a list of prices unique to each hospital and is an important indicator of payments, with the higher charges generating more revenue.[7] The chargemaster can reflect a huge mark-up of 10,000% on acetaminophen for example. In fact, almost 20% of our GDP goes to healthcare, and in comparison to other countries, our outcomes are no better.[8] In addition, prices are extremely high and can vary widely for the identical treatment; the differences between hospitals could be thousands of dollars.[9] Medicare patients and privately insured patients often never see the hospital bill and are therefore unaffected by the hospital charges. However, if higher deductibles and cost saving plans emerge in the marketplace, the portion of the bill not covered by insurance will be affected by the variation of the billed amounts or charges.[10] Uninsured and underinsured patients are most impacted by charges, because they are accountable for their bill which is based on charges. Furthermore, patients are often faced with urgent health crises, and given the variability in hospital prices[11], the challenge of being an informed consumer under such conditions is formidable.

The Septicemia cases were analyzed in two groups predicting charges and costs: “Patients that die in 1 Day” and “Patients that DO NOT die in 1 Day” resulting in 4 multiple linear regression models. Charges and costs differ by facilities adjusting for other variables. Likewise, the logistic model analyzing the probability of death of “Patients that DO NOT die in 1 day” most significant predictors are log(LOS), SOI, ROM and facility.

[6] Angus DC. 2001. Epidemiology of server sepsis in the United States: Analysis of incidence, outcome and associated costs of care. *Critical Care Medicine.* 29(7): 1303-1310.

[7] Batty M, Ippolito B. 2017. Mystery of the Chargemaster: Examining the Role of Hospital List Prices in What Patients Actually Pay. *Health Affairs*. 36(4): 689-696.

[8] Brill S. 2013. Bitter Pill: Why Medical Bills are Killing Us. *Time*. February 20, 2013; 16-55.

[9] Cooper Z, Craig S, Gaynor, Martin, & Van Rennen J. 2015. The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured. www.healthcarepricingproject.org. Accessed February 16, 2018.

[10] NYS DOH. SPARCS Hospital Inpatient Cost Transparency. 2014. https://health.data.ny.gov. Accessed February 14, 2018.

[11] Batty et al. Mystery of the Chargemaster. 689-696.

5. Conclusion

The relationship between charges and costs versus LOS, and procedures is obvious. In addition, we would intuitively expect the higher mortality risk cases to be more expensive. However, the variation in charges and costs of septicemia cases is partly explained by payer mix among other factors that differ by facility. These differences are what consumers know the least about.

Hospitals are a complex business where resources can be extremely expensive. The delicate balance of offering state-of- the-art treatment and equipment along with highly skilled physicians and nurses, attracting patients, negotiating with insurers and making a profit is complicated. Patients are interested in getting the best care, but who knew the price could be so different? Modeling of publicly reported data is very powerful; it could help guide and educate hospital administrations, policy makers and patients on how get the best value in healthcare.

6. References

Angus DC. Epidemiology of server sepsis in the United States: Analysis of incidence, outcome and associated costs of care. *Critical Care Medicine. 2001;* 29(7): 1303-1310.

Batty M, Ippolito B. Mystery of the Chargemaster: Examining the Role of Hospital List Prices in What Patients Actually Pay. *Health Affairs*, 2017; 36(4): 689-696.

Brill S. Bitter Pill: Why Medical Bills are Killing Us. *Time*. February 20, 2013; pp. 16-55.

Cooper Z, Craig S, Gaynor, Martin, & Van Rennen J. The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured. May 2015. www.healthcarepricingproject.org. Accessed February 16, 2018.

Kahn, C. Words Matter: Defining Hospital Charges, Costs and Payments - And the Numbers that Matter Mosts to Co. June 5, 2015. https://fah.org/blog/. Accessed February 14, 2018.

NYS DOH. SPARCS Hospital Inpatient Cost Transparency. January 2014. https://health.data.ny.gov. Accessed February 14, 2018.

NYSDOH Bureau of Health Informatics. SPARCS Hospital Inpatient Discharges De-identified File with Cost. January 2014. https://www.health.data.ny.gov. Accessed February 14, 2018.

NYS DOH Bureau of Primary and Acute care reimbursement. 2010 Institutional Cost Report. November 2012. https://www.health.data.ny.gov. Accessed February 14, 2018.

NYS-ITS. Data Submission Guide. June 2016. https://data.ny.gov. Accessed February 14, 2018.