Spring 2013

An Investigation into Nosocomial Infection using Data Mining

Alexander Rocke
University of New Hampshire - Main Campus, arocke1317@gmail.com

An Investigation into Nosocomial Infection using Data Mining

Abstract
The purpose of this study was to investigate potential risk factors and analyze trends that are associated with nosocomial infection using an inferential statistical methodology that would allow for the identification and future tracking of the aforementioned trends. The raw data on which statistical analyses were performed was collected and aggregated by the National Healthcare Safety Network for the year 2011 (5)

The results of the statistical analyses revealed that many factors should be taken into account when considering the causes of nosocomial infection, in paticular, that the amount of aggregated hours a certain type of ward (burn, medical, surgical, etc.) logs using invasive devices (central intravenous lines, respirators, urinary catheters) is not sufficient to explain the relative frequencies of nosocomial infections. For example, burn wards and some types of pediatric wards were particularly outstanding in their increased incidences of infection, even after the number of hours was accounted for.

A second result of the statistical analyses was the creation of intervals using a modified Poisson distribution for which a year-to-year change in the frequency of infection would be considered random. If the results of a subsequent year fell outside of that interval, the conclusion would be that a fundamental shift had occurred.

Keywords
data mining, nosocomial infection, SIR, COLSA, Molecular, Cellular and Biomedical Sciences, Biomedical Science

Subject Categories
Other Medicine and Health Sciences

An Investigation into Nosocomial Infection using Data Mining

Alexander Rocke

Table of Contents

I: Abstract

The purpose of this study was to investigate potential risk factors and analyze trends that are associated with nosocomial infection using an inferential statistical methodology that would allow for the identification and future tracking of the aforementioned trends. The raw data on which statistical analyses were performed was collected and aggregated by the National Healthcare Safety Network for the year 2011 (5)

The results of the statistical analyses revealed that many factors should be taken into account when considering the causes of nosocomial infection, in paticular, that the amount of aggregated hours a certain type of ward (burn, medical, surgical, etc.) logs using invasive devices (central intravenous lines, respirators, urinary catheters) is not sufficient to explain the relative frequencies of nosocomial infections. For example, burn wards and some types of pediatric wards were particularly outstanding in their increased incidences of infection, even after the number of hours was accounted for.

A second result of the statistical analyses was the creation of intervals using a modified Poisson distribution for which a year-to-year change in the frequency of infection would be considered random. If the results of a subsequent year fell outside of that interval, the conclusion would be that a fundamental shift had occurred.

II: Introduction

Hospital mortality is a persistent concern for the health care industry, and identifying areas of high frequency of infection is a potential method for optimizing the resources utilized to combat nosocomial mortality and infection. In the United States, millions of nosocomial infections are reported each year, with approximately 100.000 of them leading to the death of a patient. (3) The legal and financial burden on hospitals and the emotional burden on the patient and their families are proportionally enormous. In order to prevent the millions of needless infections effectively, a logical method should be used to identify the areas of the healthcare setting that contribute most, so that countermeasures might be developed efficiently. A system for tracking the year-to-year incidences of infection is also important to an undertaking such as this so that the results of any countermeasures can be observed, as well as the effects of additional unknown phenomena.

Three of the most common devices used in the healthcare industry whose use may result in infection are the central intravenous line, the ventilator, and the urinary catheter, all of which have the potential to introduce infectious agents past some of the fundamental components of the immune system, which include the skin, saliva, and immunoglobulin found on mucous membranes. As such, the use of these devices, while necessary, represents an increased risk of infection.

With the initial assumption that an equal number of hours of exposure to a central line, a ventilator or a urinary catheter should yield a proportionate number of infections, any

significant discrepancies from this distribution can be classified as non-coincidental, and may stimulate further study into the cause of those discrepancies. For example, a burn ward may have a disproportionate number of infections due to the compromised innate immune systems. The chi-square test for goodness of fit measures the probability that the observed values are accurately representing a proposed model for causes of infection. The higher the value of the chi-square test, the less likely it is that the proposed model is in fact accurate.

The SIR (standard infection ratio) is a representation of the relative number of actual infections that occurred compared to the expected number. It is calculated by dividing the observed frequency by the calculated expected value, which depends on the assumptions made. In this study, the relatively simple assumption is that the number of infections that occur should be proportional to the number of hours for which patients are attached to one of the three studied devices. Thus, a SIR of 1 implies that the number of infections that actually occurred was equal to the number of infections that were expected, where a SIR of greater than one would mean that additional infections over the expected number occurred.

The SIR should theoretically remain constant for a given ward type from year to year - however, due to essentially random fluctuations the SIR is unlikely to remain exactly the same. A confidence interval is used to determine the threshold for which we can assume that the SIR is still basically the same if the value remains within the interval, or the values for which we must believe that the SIR has actually changed due to a fundamental shift in the circumstances surrounding the ward type.

The National Healthcare Safety Network aggregates data on infections each year, so

fundamental trends in the SIRs can be identified, and closer investigation into ward types that changed significantly can be initiated.

III: Methodology

1: First, chi-square analysis was used to determine the goodness of fit for each ward type, with the assumption that the number of hours that each ward cumulatively exposes its patients to is directly proportional to the expected frequency of nosocomial infections.

2: Confidence intervals were constructed using a method described by Liddel (4),and performed by the state of New Hampshire Department of Heath and Human Services (2) which makes the assumption that the sample size is sufficiently large in order to create exact values for the upper limit for the ratio between the expected and observed frequencies. The values for the .95 confidence interval for the infection ratio are generated by the following equations:

$$[(1-(9I)^{-1}-Z(9I)^{-1/2})^3/E = \text{Lower Limit}$$

$$(1-(9(I+1))^{-1}-Z(9(I+1))^{-1/2})^3/E)] = \text{Upper Limit}$$

Where Z is equal to $3(I)^{1/2}[1-(9I)^{-1}-(I/E)^{-1/3}]$, E is the expected number of infections, and I is the observed number of infections.

IV: Results

1: Central Line IV

Ward Type	Frequency	Central Line Days	SIR	SIR L	SIR U
Outpatient Wards					
Burn	49	45,778	1.7298	0.7849	5.7610
Medical-Major teaching	148	372,229	2.3624	0.9557	1.8473
Medical -All other	298	384,630	0.9013	0.9479	1.0534
Medical cardiac	293	376,962	0.9094	0.9473	1.0792
-Major teaching	192	417,461	1.5948	0.9534	1.1603
Medical/surgical -≤15 beds	1170	841,016	0.4138	0.9727	0.7090
Medical/surgical > 15 beds	518	1,177,318	1.2538	0.9799	0.6471
Neurologic	24	37,952	0.9712	0.5212	0.8183
Neurosurgical	95	154,375	1.1543	0.8749	1.0907
Pediatric cardiothoracic	30	79,803	2.9313	0.8230	2.4560
Pediatric medical	24	23,730	0.9933	0.4656	2.0951
Pediatric medical/surgical	204	270,003	1.2387	0.9391	1.8354
Respiratory	7	10,760	0.4541	-0.0018	0.1524
Surgical-Major teaching	127	297,551	1.7102	0.9349	1.1497
Surgical -All other	149	227,644	0.8035	0.8956	0.5841
Surgical cardiothoracic	294	554,719	0.9352	0.9545	0.5114
Trauma	94	197,290	2.1303	0.9215	2.1237
Chi-Square Value	**559.15**				
Inpatient Wards					
Acute stroke	1	8,545	0.0830	-38.8067	0.0007
Behavioral health/psychiatry	4	3,360	0.0766	-2.6032	1.5767
Burn	5	2,028	1.2449	-2.8784	5.3012
Genitourinary	11	15,947	1.2449	0.0219	0.4985
Geronotology	5	4,320	1.0374	-0.6879	1.4777
Gynecology	7	10,052	0.2421	-0.0025	0.5203
Jail	22	7,899	2.7388	0.0776	6.2536
Medical	665	615,168	1.5417	0.9591	1.2270
Medical/surgical	1,209	1,304,991	1.0034	0.9785	0.9034
Neurologic	25	32,628	0.798	0.3032	0.6154
Neurosurgical	25	31,344	0.798	0.2973	0.6709
Orthopedic	52	92,107	0.3786	0.6305	0.3003
Orthopedic trauma	21	13,028	2.6143	0.1245	2.5722
Pediatric medical	40	33,086	1.7171	0.4411	1.5278
Pediatric medical/surgical	156	105,530	1.0385	0.8162	2.1906
Pediatric orthopedic	2	1,242	0.4150	-77.0149	2.8166
Pediatric rehabilitation	12	4,283	1.2449	-0.0002	6.3457
Pediatric surgical	18	13,765	2.8011	0.0944	1.7798
Pulmonary	36	34,422	2.3588	0.4172	1.1601
Rehabilitation	53	98,631	0.0273	0.6415	0.2651
Surgical	305	303,472	1.2871	0.9142	1.0646
Vascular Surgery	16	21,217	0.9959	0.1176	0.5999
Chi-Square Value	**147.46**				

2: Urinary Catheter

	Frequency	Catheter Days	SIR	SIR L	SIR U
Critical care units					
Burn	23	24,324	1.3287	0.7201	4.7717
Medical -Major teaching	67	192,002	1.8642	0.9401	1.5386
Medical -All other	110	232,454	1.0533	0.9409	0.9169
Medical cardiac	139	213,535	0.7915	0.9372	0.9803
Medical/Surgical -Major teaching	98	263,186	1.5917	0.9533	1.2901
Medical/Surgical -All other, ≤15 beds	397	434,729	0.3715	0.9589	0.3958
Medical/Surgical -All other, >15 beds	201	596,233	1.0180	0.9702	0.4066
Neurologic	12	27,681	1.8602	0.6758	2.2758
Neurosurgical	45	110,797	2.6338	0.9259	3.6730
Pediatric cardiothoracic	10	8,988	0.5581	0.1674	1.4281
Pediatric medical	6	1,527	0.2657	-0.6548	3.6404
Pediatric medical/surgical	78	57,420	0.4327	0.7969	1.2718
Surgical -Major teaching	59	157,384	2.1214	0.9362	2.2157
Surgical -All other	53	118,919	0.9125	0.8710	0.5984
Surgical cardiothoracic	124	239,246	0.7951	0.9350	0.6151
Trauma	51	151,217	2.5428	0.9368	2.5283
Chi Square Value:	**873.25**				
Inpatient					
Acute stroke	8	8896	0.1874	-2.4814	0.0016
Behavioral health/psychiatry	79	6687	0.1234	0.0169	1.6523
Genitourinary	7	10684	1.0709	0.0143	0.3665
Gerontology	5	3216	0.5997	-1.2651	0.7303
Gynecology	22	17307	0.3407	0.0448	0.0993
Labor and delivery	34	11051	0.1323	-0.0001	0.0859
Labor, delivery, recovery, postpartum suite	7	22853	1.7135	0.1855	0.1704
Medical	341	333155	1.1849	0.9503	1.1418
Medical/Surgical	877	854649	1.0719	0.9791	0.9399
Neurologic	23	25030	1.5971	0.5165	1.6485
Neurosurgical	22	34773	2.5897	0.6535	2.0100
Orthopedic	102	127082	1.1980	0.8530	0.7130
Orthopedic trauma	5	8138	0.8996	-0.0045	0.2088
Pediatric medical/surgical	84	13283	0.1606	0.1439	0.8145
Pediatric medical	14	1379	0.1071	-38.6762	1.0688
Postpartum	94	49862	0.1595	0.3648	0.0247
Pulmonary	11	14676	1.2948	0.1718	0.7400
Rehabilitation	181	78514	1.0313	0.8691	3.7944
Pediatric rehabilitation	7	723	0.2142	-87.1800	3.3240
Surgical	170	233119	1.5963	0.9275	1.0535
Telemetry	32	20841	0.2577	0.0802	0.0736
Vascular surgery	8	8324	1.9678	0.1236	2.6064
Chi-Square Value	**269.3**				

3: Ventilator

Critical Care	**Frequency**	**Ventilator**	**SIR**	**SIR L**	**SIR U**
Critical care units		**Days**			
Burn	24	15,379	3.1774	1.5465	4.3992
Medical-Major teaching	78	153,408	0.7444	0.6174	0.7860
Medical -All other	106	132,014	0.5490	0.4218	0.5785
Medical cardiac	125	103,375	0.6958	0.5261	0.7426
Medical-Major teaching	101	167,857	1.0042	0.8653	1.0651
Medical/surgical ≤15 beds	359	221,857	0.6335	0.5460	0.6606
Medical/surgical -All other >15 beds	209	358,913	0.5920	0.5348	0.6110
Neurologic	15	14,837	2.6274	1.1962	3.5625
Neurosurgical	46	53,966	1.6787	1.2315	1.9149
Pediatric cardiothoracic	13	26,784	0.4100	0.0575	0.4629
Pediatric medical	10	8,737	0.6284	0.0000	0.8072
Pediatric medical/surgical	82	103,094	0.6391	0.4746	0.6807
Respiratory	6	6,659	0.0000		
Surgical -Major teaching	70	106,736	1.9238	1.6121	2.1252
Surgical -All other	61	71,746	1.3851	1.0644	1.5380
Surgical cardiothoracic	123	132,307	0.9047	0.7473	0.9641
Trauma	48	92,460	3.2957	2.7126	3.7828
Chi Squared Value	**2957.36**				

V: Discussion

The major findings from this analysis indicate that the model used (where the number of infections is proportional to the time factor) is not sufficient to explain the pattern of observed infections. The major example of this result can be seen in the statistics for the burn ward, in which the SIR is significantly higher than 1 in all cases. An easy rationale for this result would be the immunocompromised state that many burn patients experience, however, additional study would be indicated to conclusively confirm this. Potential avenues of further analysis in this regard would be to collect non-aggregated data so that wards could be classified on different criteria, such as the accumulated experience of the medical staff, the location of the ward and the size of the ward. For an even further, more rigorous analysis, individual cases of infection could be examined, so that immunocompromised states, pre-existing conditions, time of day, time of year, microorganism involved, antibiotic use, or other circumstances could be taken into consideration, greatly strengthening the value of the conclusions that can be drawn from the analyses.

Another finding from the analyses for urinary catheters and intravenous lines indicate that the SIR for outpatient wards is slightly elevated from 1.0 in inpatient wards and is decreased for critical care wards for these devices, even after taking into account the difference in patient hours on the invasive device. One potential explanation for this phenomenon is that in critical care wards, there are a greater number of patients overall, but with fewer hours on a device per patient. In this case, the implication would be that the risk of infection goes up the longer a single patient is attached to a urinary catheter or an

intravenous line. To confirm this hypothesis, the next step would be to analyze individual cases with infections, and determine the time after initial attachment to an invasive device it takes for an individual patient to develop an infection. Should the results of this next study reveal that there is a statistical bias towards longer or shorter attachments to devices, a potential corrective step would be to analyze the probability of infection based on the number of patients treated in each type of ward, rather than on the basis of total hours. It was also found that the variance for each ward type is significantly higher on average for critical care wards than for inpatient wards (as is reflected in the chi-squared value), so additional years of data would be useful in determining if the data for 2011 is indicative of a larger pattern.

Additionally, the major contributor (the ward type with the most number of pooled hours) consistently throughout each analysis had an SIR of less than or approximately equal to 1.0. This implies that the wards with the most number of hours do not account for an increased risk of infection as much as wards with a lower number of hours. One potential explanation for this would be that wards with a large number of pooled patient-device interaction hours have some quality about them that correlates to a lower probability of infection. This could be an intentional effect, such as more caution against infection in wards such as these, or it would be a obscured effect that may warrant further study. One way to study this would be to take note of the conditions of these types of wards, separate them further based on potential factors, and analyze the rate of infection based on these groupings. One way to group them would be by the level of stringency of anti-infection procedures, for example, how many times the staff washes their hands when interacting with the devices at those types of wards as opposed to others which make up a less significant proportion of the pooled patient-device hours.

The confidence intervals found for SIRs can be easily converted to intervals for frequencies of infections by multiplying the SIR bounds by an expected number of infections. With either of these interval types, the trends in infection incidences can be monitored from year to year by performing a similar analysis on data from each year. One caveat of the method used is that the confidence interval becomes much larger and thus more unreliable as the frequencies of infection decrease. Therefore, even though the intervals are calculated for each ward type, analyses for those entries that have low numbers of infections should be treated as somewhat unreliable. In some cases, this can be remedied by combining ward categories which are very similar, in order to create a broader conclusion. Again, with non-aggregated data, the possibility for large amounts of differentiation and classification is high; that will greatly improve the versatility of the analyses.

Another potential application of this data would be to perform an analysis similar to that performed by Hautemanière, in order to analyze the role of these infections in mortality in the healthcare setting (1). This would require a case-by-case breakdown of the data, as described.

VI: Conclusion

The judicious application of statistical analysis has a large role to play in the intelligent and focused process of improving healthcare in the modern era as the increasingly large scale of the industry continues to obfuscate the chains of causality. In a field in which the slightest of errors is translated into serious emotional and legal consequences, and a relative decrease in expendable resources, the ability to identify problem areas is paramount for the cost-effective regulation of nosocomial infection. The results of the analyses are only examples of the hypotheses that can be extracted and subsequently tested, and the addition of more information to the analyses will only strengthen the resulting hypotheses, which, with the systematic use of additional analysis on a more differentiated level will reveal accurate trends in the data which can be used to guide the creation of healthcare policy to most effectively meet the needs of the industry and the consumer.

VII: Literature Cited

1: Hautemanière, et al Identifying possible deaths associated with nosocomial infection in a hospital by data mining, American Journal of Infection Control, Volume 39, Issue 2, March 2011, Pages 118-122, ISSN 0196-6553, 10.1016/j.ajic.2010.04.216.

2: Heathcare Related Infections, 2011 Report State of New Hampshire Department of Health and Human Services, March 1st 2013

3: Klevens et. al. Estimating Health Care-Associated Infections and Deaths in U.S. Hospitals, 2002 ,http://www.cdc.gov/HAI/pdfs/hai/infections_deaths.pdf

4:Liddel, "Simple Exact Analysis of the Standardized Mortality Ratio", *Journal of Epidemiology and Community Health 1984-38*

*5:*National Healthcare Safety Network 2011 Annual Report
"http://www.cdc.gov/nhsn/dataStat.html#ar"