TEMPLATE MATCHING FOR HOSPITAL COMPARISON: AN APPLICATION TO BIRTH EVENT DATA IN ITALY

Massimo Cannas [1], Paolo Berta [2] and Francesco Mola [1]

[1] Department of Economics and Management, University of Cagliari, (e-mail: massimo.cannas@unica.it)

[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, (e-mail: paolo.berta@unimib.it)

ABSTRACT: Quality evaluation in healthcare has obtained a growing attention in the statistical literature. In order to evaluate hospital performances by comparing hospital outcomes it is necessary to remove the bias due the different case-mix in each hospital, which is usually done using statistical modelling as a risk adjustment tool. In this paper we standardize the patients' allocation to each hospital adopting a template matching approach. We illustrate the method by comparing quality indicators of hospital deliveries obtained from official birth event abstracts, both within and between the hospitals of Sardinia and Lombardy regions (Italy). The results suggest that this method allows a fair comparison among providers, making easier for the policy makers the identification of potential distortions in patients' selection and outcomes.

KEYWORDS: case-mix, hospital evaluation, template matching, multilevel model.

1 Introduction

In the last decade, quality evaluation in healthcare has obtained a wide attention in the statistical literature. One of the most important point to take care comparing the hospitals on their outcomes' performance is the risk-adjustment model applied to standardize the different patients' severity in each hospital. Furthermore, the implications connected with the hierarchical structure of the data (patients nested within hospitals) moved the literature on this topic to the adoption of the multilevel modelling, in order to disentangle the heterogeneity at the hospital and patient level.

In the hospital evaluation framework, the allocation of the patients to the hospitals can be figured as a group of subjects in a multi-treatment setting, where the treatments are the hospitals. Because the patients are not randomly assigned to the hospitals, a bias selection problem affects every treatment effect analysis. To avoid this problem Silber *et al.* , 2014 proposed a so-called *template matching* approach, by which only average patients treated by quite all the hospitals are involved in the evaluation process. Possible significant differences on these average patients across the hospitals either prompt for explanations or can be readily interpreted as quality indicators when observed on final outcomes (e.g. mortality rates). This new approach is here tested for the first time to compare the incidence of several birth outcomes within and between the hospitals of two Italians regions. We also consider template matching integration with a classic evaluation approach based on a multilevel model.

2 Data and Methods

Template matching method is based on the ”fair exam” idea (Silber *et al.* , 2014). We define ”fair” an exam that asks the students the same questions. Translating this concept to the healthcare sector, we have to think of an hospital performance evaluation like an exam where the questions (the patients) asked to the students (the hospitals) should be on average the same. This is not usually the case because of the differences in the capabilities and backgrounds of the patients across the hospitals. But this means that a performance evaluation can be defined as fair if a template of similar patients could be found in each of the hospitals under evaluation. Following Silber *et al.* , 2014 we applied classic and Coarsened Exact Matching (CEM, Iacus *et al.* , 2011) to perform a multivariate matching on expectant mothers' covariates. The result of the matching is a selection of similar women in each of the hospitals. The selected subsets are then compared using statistical tests on C-section rates and other quality indicators, both across and within the hospitals of Sardinia and Lombardy.

In order to define the template we used mother's age, gestational age, nulliparity condition, plus infant's weight and a set of indicators for morbidities and the frequency of control checks (e.g. sonograms) during pregnancy. We identified the template as the sample with lower multivariate distance from the overall means for the variables above in the entire population. In this short version of the paper we concentrate the analysis within Sardinia, for which we used a template size of 100, which makes possible a matching ratio of 2 sacrificing only smallest hospitals and leaving a total of 17 hospitals to be compared. In the second step we used CEM and other matching algorithms to match the patients of each hospital with the template patients. In this way we extracted n subsets with almost the same size of the template, where $n = 17$ is the number of hospitals included in the analysis.

The final dataset is then composed by $100 \times k$ observations, $k \leq n$ being the number of hospitals that matched with the template. In fact, it could be that some hospitals admits women with very specific characteristics so that, based on the template concept, these hospitals cannot take the "fair exam". Note that these hospitals are different in terms of case-mix, and for the policy makers, it would be interesting to understand whether this difference is justified or whether is the result of an improper patients' selection. The matching procedure adjust the differences in terms of case-mix among hospitals, and this allow us to compare an hospital performance in terms of Apgar index and C-section and labor induction rates using simple statistical tests for hospital mean differences.

The individual data analyzed in this paper are gathered from Birth Assistance Certificate (CEDAP) occurring in Sardinia and Lombardy Region from 2011 to 2012. CEDAP data are a useful tool to detect and identify critical aspects in the process of care during pregnancy and childbirth and to make comparative analyzes of birth centers. CEDAP contains detailed information about socio-demographic status of mother and father, the pregnancy and newborn's status.

3 Results

The chosen template was matched with women in all hospitals with at least 200 observations. The matching variables were the same used in template construction but different calipers were used for some of them in order to calibrate a very tight matching for those variables affecting the outcomes in a particularly strong way. For example we matched exactly on fetus presentation, so that the proportion of presentations are exactly the same in the template and in the matched subsets. We used a large caliper of two standard deviation to exclude only major outliers for all other variables. More than 90 observations matched the template in each of the hospitals, with the exception of hospital 8, were only 15 observations matched. This is because women delivering in this hospital were mostly primiparae and exposed to far less checks than women in the template. The template and matched variables are compared in Table 1 and turn out to be very similar.

Delivery outcomes for each hospital are compared with outcomes from all remaining hospitals in Table 2. That is, the performance of each hospital is evaluated by comparing the average value of the outcome for women delivering in that hospital with the average value for women delivering in all other hospitals. Significant difference can be observed in Apgar scores of hospitals number 1,2,3,4,5 and 10 which are significantly different from scores outside each of these hospitals. Also, hospitals 6 and 8 resorted to caesarean section significantly less than the rest of the hospitals. Finally, labour induction use strongly varies across hospitals with several comparison resulting in a rejection of the null hypothesis of no difference. Since these differences are obtained *ceteris paribus* on the common ground of the template they provide a reasonable input for policy makers.

4 Conclusion

These preliminary results support the hypothesis that the usual performance evaluation approach could suffer from a strong bias selection due to the not random allocation of the patients to the hospitals and due to the cream skimming of the patients performed by some hospitals. The approach presented in this paper illustrates the use of template matching in the performance evaluation of the birth outcomes. The results demonstrate that template matching helps excluding from the performance evaluation analysis some hospitals that cannot be compared because of their specific case-mix. Moreover, the results of the hospital comparisons can be considered most effective because the hospitals are subjected to a ”fair exam”. The final version of this short paper will be improved adding the comparison among the Lombardy data and the comparison between the two regions.

References

GOLDSTEIN, HARVEY, & SPIEGELHALTER, DAVID J. 1996. League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 385–443.

IACUS, STEFANO M, KING, GARY, & PORRO, GIUSEPPE. 2011. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, **106**(493), 345–361.

LEYLAND, ALASTAIR H, & GOLDSTEIN, HARVEY. 2001. *Multilevel modelling of health statistics*. Wiley.

SILBER, JEFFREY H, ROSENBAUM, PAUL R, ROSS, RICHARD N, LUDWIG, JUSTIN M, WANG, WEI, NIKNAM, BIJAN A, MUKHERJEE, NABANITA, SAYNISCH, PHILIP A, EVEN-SHOSHAN, ORIT, KELZ, RACHEL R, *et al.* . 2014. Template matching for auditing hospital cost and quality. *Health services research*, **49**(5), 1446–1474.

Table 1. *Comparing template with hospital matched subsets.*

Hospital		Age	Gest	Decorso	Length	Weight	N. of	N. of	N. of
matched			age		(grams)	(cm)	prev. del.	prev. cs	sonograms
template	mean	32.79	38.73	0.09	49.25	3169.00	0.48	0.10	6.26
	sd	5.68	2.28	0.29	2.98	522.13	0.80	0.41	2.42
1	mean	32.74	38.58	0.08	49.01	3142.78	0.41	0.07	6.22
	pvalue	0.952	0.630	0.852	0.563	0.720	0.520	0.588	0.901
2	mean	33.00	38.76	0.08	49.25	3141.92	0.44	0.07	6.33
	pvalue	0.793	0.929	0.818	0.995	0.715	0.741	0.566	0.829
3	mean	32.85	38.95	0.08	49.06	3153.28	0.46	0.07	6.39
	pvalue	0.938	0.477	0.818	0.633	0.828	0.887	0.566	0.681
4	mean	32.76	38.97	0.07	49.42	3177.04	0.45	0.07	6.23
	pvalue	0.970	0.386	0.633	0.640	0.909	0.774	0.577	0.941
5	mean	33.09	38.89	0.07	49.37	3182.35	0.45	0.07	6.31
	pvalue	0.702	0.584	0.633	0.746	0.847	0.774	0.577	0.891
6	mean	32.70	39.04	0.07	49.34	3173.27	0.45	0.07	6.35
	pvalue	0.909	0.279	0.633	0.810	0.952	0.780	0.577	0.796
7	mean	32.78	38.96	0.07	49.63	3177.76	0.46	0.07	6.28
	pvalue	0.991	0.417	0.633	0.292	0.902	0.852	0.577	0.963
8	mean	31.32	39.18	0.00	49.91	3167.73	0.09	0.00	1.09
	pvalue	0.327	0.340	0.002	0.226	0.988	0.003	0.018	0.000
9	mean	32.89	39.09	0.07	49.48	3196.17	0.46	0.07	6.16
	pvalue	0.902	0.192	0.633	0.505	0.691	0.851	0.577	0.775
10	mean	32.73	39.08	0.06	49.63	3171.19	0.47	0.07	6.55
	pvalue	0.935	0.207	0.458	0.290	0.975	0.959	0.588	0.374
11	mean	32.84	38.97	0.07	49.24	3200.82	0.43	0.07	6.43
	pvalue	0.948	0.371	0.648	0.971	0.637	0.664	0.588	0.605
12	mean	33.18	39.03	0.07	49.46	3203.93	0.51	0.10	6.44
	pvalue	0.615	0.292	0.633	0.559	0.621	0.792	0.973	0.574
13	mean	32.48	38.97	0.07	49.68	3192.01	0.46	0.07	6.62
	pvalue	0.683	0.387	0.648	0.230	0.739	0.882	0.588	0.262
14	mean	32.86	39.00	0.07	49.37	3192.26	0.41	0.07	6.30
	pvalue	0.933	0.323	0.633	0.744	0.736	0.497	0.577	0.912
15	mean	32.41	38.94	0.07	49.11	3236.43	0.43	0.07	6.36
	pvalue	0.626	0.459	0.633	0.689	0.315	0.631	0.577	0.762
16	mean	33.53	39.20	0.02	49.48	3182.42	0.48	0.10	6.93
	pvalue	0.351	0.079	0.039	0.507	0.847	0.974	0.984	0.029
17	mean	32.16	39.25	0.02	49.47	3226.43	0.44	0.08	6.91
	pvalue	0.418	0.052	0.039	0.547	0.396	0.698	0.660	0.033

Table 2. *Comparison of outcomes between an hospital matched subset and all matched subsets from other hospitals*

Hospital		mean hosp[i]	mean hosp[-i]	mean difference	pvalue
1	Apgar 5 min	9.02	9.69	-0.67	0.00
	C Section	0.43	0.35	0.08	0.11
	Labor induction	0.14	0.21	-0.07	0.07
2	Apgar 5 min	9.79	9.64	0.15	0.02
	C Section	0.32	0.36	-0.03	0.51
	Labor induction	0.21	0.21	0.00	0.92
3	Apgar 5 min	9.92	9.63	0.29	0.00
	C Section	0.39	0.35	0.04	0.40
	Labor induction	0.43	0.19	0.24	0.00
4	Apgar 5 min	9.72	9.64	0.08	0.18
	C Section	0.37	0.35	0.01	0.78
	Labor induction	0.21	0.21	0.01	0.88
5	Apgar 5 min	9.36	9.67	-0.31	0.02
	C Section	0.38	0.35	0.03	0.62
	Labor induction	0.19	0.21	-0.01	0.72
6	Apgar 5 min	9.89	9.63	0.26	0.00
	C Section	0.22	0.36	-0.14	0.00
	Labor induction	0.18	0.21	-0.03	0.53
7	Apgar 5 min	9.72	9.64	0.08	0.30
	C Section	0.27	0.36	-0.09	0.04
	Labor induction	0.21	0.21	0.01	0.88
8	Apgar 5 min	9.00	9.65	-0.65	0.00
	C Section	0.27	0.35	-0.08	0.58
	Labor induction	0.00	0.21	-0.21	0.00
9	Apgar 5 min	9.92	9.63	0.29	0.00
	C Section	0.32	0.36	-0.04	0.41
	Labor induction	0.17	0.21	-0.04	0.36
10	Apgar 5 min	9.78	9.64	0.14	0.03
	C Section	0.32	0.36	-0.04	0.46
	Labor induction	0.21	0.21	-0.00	0.97
11	Apgar 5 min	9.75	9.64	0.11	0.06
	C Section	0.39	0.35	0.04	0.43
	Labor induction	0.31	0.20	0.11	0.03
12	Apgar 5 min	9.68	9.65	0.04	0.55
	C Section	0.34	0.35	-0.02	0.71
	Labor induction	0.23	0.21	0.03	0.52
13	Apgar 5 min	9.59	9.65	-0.06	0.37
	C Section	0.37	0.35	0.02	0.72
	Labor induction	0.22	0.21	0.01	0.83
14	Apgar 5 min	9.32	9.67	-0.35	0.00
	C Section	0.28	0.36	-0.08	0.08
	Labor induction	0.11	0.21	-0.10	0.00
15	Apgar 5 min	9.58	9.65	-0.07	0.39
	C Section	0.42	0.35	0.07	0.18
	Labor induction	0.15	0.21	-0.06	0.13
16	Apgar 5 min	9.87	9.63	0.23	0.04
	C Section	0.53	0.34	0.18	0.00
	Labor induction	0.13	0.21	-0.08	0.03
17	Apgar 5 min	9.53	9.66	-0.13	0.03
	C Section	0.34	0.35	-0.01	0.79
	Labor induction	0.21	0.21	0.00	0.98