Nursing Informatics 2016
W. Sermeus et al. (Eds.)

doi:10.3233/978-1-61499-658-3-402

Identifying Outliers in Data from Patient Record

Dieter BAUMBERGER[a,1] and Reto BUERGIN[a]
[a]*Research and Development LEP AG, St.Gallen; Switzerland*

Abstract. It is important for health services to be able to identify potential outliers with minimal effort as part of their daily evaluation of care data from patient record. This study evaluates the suitability of three statistical methods for identifying nursing outliers. The results show that by using methods implemented in the nursing workload measurement system "LEP" with reference to real data, unusual LEP minute profiles (movement, nutrition and so on) can be identified with little effort and therefore seem promising for application to the health services' daily evaluation process. The lessons learned are used to create requirement criteria for the further development of software solutions. It is recommended that the methods for identifying outliers in the daily evaluation process should be standardized in order to increase the efficiency of secondary use of care data from patient record.

Keywords. Secondary use of Care Data, Evaluation of Care Data, Identification of Outliers, Statistical Methods, Diagnosis Related Groups, Nursing Workload. Electronic Patient Records System.

1. Introduction

Electronic patient record is being used increasingly often by today's health services as a reference point for the secondary use of nursing data [1, 2]. Around 250 health services in Germany, Switzerland, Austria and Italy are using LEP ("Leistungserfassung in der Pflege", nursing workload measurement) for the statistical evaluation of nursing data [3]. The latest generation, LEP Nursing 3, is a classification for around 500 nursing interventions that will be employed in electronic patient record and for automated nursing workload measurement. [4, 5]. It is semantically focused on ICNP and is structured in line with ISO-18104 [6, 7]. When applied in nursing practice, it follows the methodical approach of 'collect once, use in many cases', with regard to documented data [8]. LEP service data is automatically channeled from patient record in order to perform statistical evaluations. This may be to gather evidence of the treatment quality, for analysis of the nursing workload, for benchmarking, the calculation of nursing costs or DRG coding. By using this method, any additional measurement expense required for performing statistical evaluations will be limited to the essential. 'Bureaucratic burdens' and redundant data collection at the point of care may be avoided [9]. As part of the daily evaluation process, it is important for the health services to be able to identify potential outliers in the care data from patient record with minimal effort. A health service may be interested in methods for identifying outliers for various reasons. For example, the objective of identifying

[1] Corresponding Author: Dieter Baumberger, RN, NEd, MScN, PhD, Research and Development LEP AG, Blarerstr. 7, CH- 9000 St.Gallen, Switzerland. E-mail: dieter.baumberger@lep.ch

outliers may simply be to ascertain the volume of documentation or coding errors, but it may also be to determine medicinal or nursing outliers.

The driving force behind this study was enabling the discovery of new information in relation to care data from patient record. The aim was to find statistical methods to identify nursing outliers, which could be put to use in the health services' day-to-day evaluation process. The health services would then be able use the subsequent knowledge of outliers to gain new insights towards the improvement of treatment processes or for change management.

2. Methods

The statistical methods 'k-means' [10], 'k-medoids' [11] and 'trimmed k-means' [12] were implemented for the identification of outliers and evaluated using real data from health services. We are specifically interested in the extent to which unusual LEP minute profiles comprising several dimensions (movement, nutrition and so on) can be identified using these methods. The three methods extract standard LEP minute profiles from the given data, and individual profiles deviating substantially from these standard profiles are classified as outliers.

Table 1 shows six artificial LEP minute profiles. The first three are identical, thus their pattern may be seen as a standard profile. The profiles 4 to 6 show different outlier patterns. In profile 4 the number of LEP minutes in the activity dimension is substantially higher than compared to other, profile 5 has a somewhat inverse pattern compared to profiles 1 to 3 and in profile 6 all entries are zero.

Table 1. Artificial data of LEP-minutes profiles of six persons. Each profile has four values in minutes concerning the LEP dimensions movement, activity, safety and medication.

Person	Movement	Activity	Safety	Medication
1	10	30	10	30
2	10	30	10	30
3	10	30	10	30
4	10	150	10	30
5	30	10	30	10
6	0	0	0	0

Table 2 shows the resulting standard profiles from using the three considered methods. For example the standard profile from k-mediods is identical to the pattern of the profiles 1 to 3 of **Table 1**. Based on the standard profiles it is possible measure the extent to which individual profiles are outliers. This extent may be quantified by the Euclidean distance between the standard profiles and the individual profiles.

Table 2. Extracted standard profiles using the methods k-means, k-mediods and trimmed k-means for the data of table 1.

	Movement	Activity	Safety	Medication
k-means	11.6	41.6	11.6	21.6
k-mediods	10	30	10	30
trimmed k-means	12	20	12	20

Table 3 gives the Euclidean distances between the individual profiles of **Table 1** and the standard profiles of **Table 2**. With all three methods the profiles 4 to 6 deviate most. Profile 4 is best identified by the trimmed k-means method, and k-means detects best the profiles of person 5 and person 6.

Table 3. Euclidean distances between individual profiles of table 1 and the standard profiles of table 2.

Person	k-means	k-mediods	trimmed k-means
1	14.5	0	14.4
2	14.5	0	14.4
3	14.5	0	14.4
4	108.7	120	130.4
5	42.6	40	29.1
6	49.8	44.7	33

In the applications to real data below the three methods were applied separately for DRGs (Diagnoses Related Groups, such as strokes for example), so that outliers can be defined specifically according to patient-groups. Data transformations, such as logarithms, were examined for the minute scale. The data transformations appear to be encouraging for certain types of outliers. The three methods, 'k-means', 'k-medoid' and 'trimmed k-means', are freely available in the statistical software environment R [13].

3. Results

The three methods were evaluated by using real care data from twelve Swiss hospitals. The used data include 73’930 LEP minutes profiles from 213 DRGs [14]. Each LEP minutes profile includes LEP minutes of 15 dimensions.

The methods were applied separately for each of the 213 DRGs. Logarithm transformations were used because these improved the identification of profiles with unusually low LEP minutes.

We found that for most of the DRGs the methods identify very similar or exactly the same outliers. Figure 1 shows with B61Z[2] one of the rare DRGs for which the outlier identification differs between the three methods. In this case, k-means and trimmed k-means identify both profiles with low and high LEP minutes while the k-mediods method highlights solely profiles with low minutes.

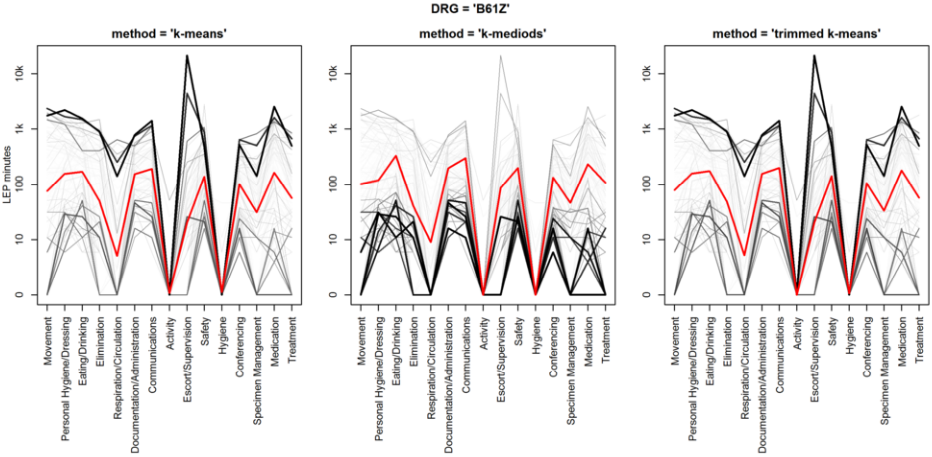

Figure 1. Outlier detection for the DRG B61Z (N=91). Red lines present extracted standard LEP minutes profiles. Grey lines present observed LEP minute profiles. Line widths and grey levels are proportional to the Euclidean distance of the profiles to the standard profile.

[2] SwissDRG B61Z: Specific acute disorders and injuries of the spinal cord (Original description: Bestimmte akute Erkrankungen und Verletzungen des Rückenmarks)

The implementation proved to be particularly useful for identifying outliers, where the documented LEP minutes were very high in particular dimensions. **Figure 2** exemplifies this by using the results for the DRG I05Z[3]. In this case, the results of the three methods do not drastically differ.

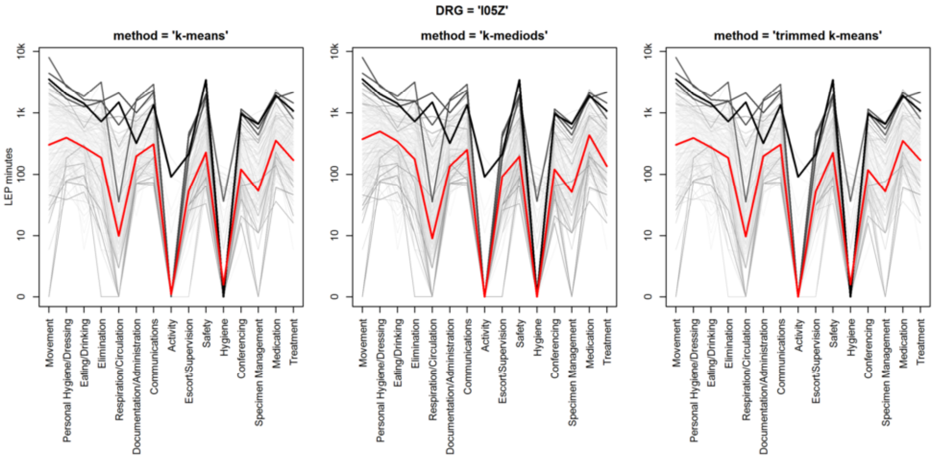

Figure 2. Outlier detection for the DRG I05Z (N=291). Red lines present extracted standard LEP minute profiles and grey lines observed LEP minutes profiles.

In practice the identified outliers may be classified. Some outliers may be coding errors, some may not be outliers from a medical perspective but show a rare pattern and some outliers may be studied in more detail.

The lessons learned make it clear that the implementation of these methods into the health service evaluation processes can provide valuable and sometimes surprising insights into the data and can help to improve the data quality. However, we found that the use of these methods requires a certain degree of knowledge of statistical methods so that interesting outlier patterns can be identified in a reliable manner.

4. Discussion

The results show that outliers can easily be identified from the care data from patient record using the three statistical methods. The knowledge gleaned from the identified outliers is useful for change management within the health services. Lessons learned from the implementation of statistical methods have been used for generating requirement criteria for software solutions for identifying outliers in LEP service data from patient record; the fact that scatter diagrams need to be available, for example. It is recommended that the methods for identifying outliers in the health services' daily evaluation process should be standardized in order to increase the efficiency of the secondary use of care data from patient record.

---

[3] DRG I05Z Other major joint replacement or revision or hip joint replacement without complicated diagnosis, without arthrodesis, without complex procedure, with extremely severe CC (original description: Anderer grosser Gelenkersatz oder Revision oder Ersatz des Hüftgelenkes ohne komplizierende Diagnose, ohne Arthrodese, ohne komplexen Eingriff, mit äusserst schweren CC)

References

[1] W.O. Hackl, F. Rauchegger, E. Ammenwerth, A Nursing Intelligence System to Support Secondary Use of Nursing Routine Data. *Applied clinical informatics* **6** (2015), 418–428.

[2] B. Sellemann, J. Stausberg, U. Hübner, Nursing routine data as a basis for association analysis in the domain of nursing knowledge. *Nursing informatics* (2012), 366.

[3] LEP AG. References. LEP Method for documentation and analysis of care activities. http://lep.ch/de/referenzen.html, last access: 17.09.2015.

[4] D. Baumberger, G. Kühne, LEP Nursing 3 for the linkage of electronic patient record and nursing workload measurement. In: N. Oud, F. Sheerin, M. Ehnfors, W. Sermeus, (ed.), *Proceedings of the 6th European Conference of ACENDIO. Nursing Communication in Multidisciplinary Practice* (161), Oud Consultancy, Amsterdam, 2007.

[5] M. Oertle, A. Baumgartner, Ausmass der vollständigen Automatisierung von Pflegeleistungserfassungen basierend auf der elektronischen Patientenakte. *Swiss Medical Informatics* **69** (2010), 37–40.

[6] D. Baumberger, K. Jansen, N.R. Hardiker, M. Studer, P. Tackenberg, P. König, ICNP Projects. Mapping ICNP and LEP Nursing. http://www.icn.ch/details/2/154.html [15.09.2015].

[7] D. Baumberger, M. Stadler, D. Buchmann, Mapping LEP Nursing 3 and ISO reference terminology model for nursing action (ISO-RTM). *Swiss Medical Informatics* **28** (2012), 47–48.

[8] J.J. Cimino, Collect once, use many. Enabling the reuse of clinical data through controlled terminologies. *J AHIMA* **78** (2007), 24-9.

[9] P. Schulz, Kontroversen in der Medizinischen Informatik. Wozu benötigen wir standardisierte Terminologien wie SNOMED CT? *Swiss Medical Informatics* **73**, (2011), 27–32.

[10] J.A. Hartigan, M.A. Wong, A K-Means clustering algorithm. *Applied Statistics* **28** (1979), 100-108.

[11] A. Reynolds, G. Richards, B. de la Iglesia, V. Rayward-Smith, Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, (1992), 475–504.

[12] J.A. Cuesta-Albertos, A. Gordaliza, C. Matrán, Trimmed k-means: an attempt to robustify quantizers. *Annals of Statistics* (1997), 553-576.

[13] R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, 2015.

[14] D. Baumberger, W. Bosshard, L. Portenier, M. Wittwer, R. Bürgin, M. Müller, Projekt SwissDRG und Pflege. Teilaufgabe 1. Quantitative Analyse Pflegeaufwand. http://www.nursingdrg.ch/files/4714/2798/8923/SwissDRG_Pflege_Analyse_Pflegeaufwand_2009_10_14.pdf, last access: 17.09.2015.