A SYSTEMATIC REVIEW OF OUTLIERS DETECTION TECHNIQUES IN MEDICAL DATA: PRELIMINARY STUDY

Juliano Gaspar\textsuperscript{1,2}, Emanuel Catumbela\textsuperscript{1,2}, Bernardo Marques\textsuperscript{1,2}

\textsuperscript{1} Department of Biostatistics and Medical Informatics, Faculty of Medicine, University of Porto, Portugal

\textsuperscript{2} CINTESIS - Center for Research in Health Technologies and Information Systems, University of Porto, Portugal

jgaspar@med.up.pt, ecatumbela@med.up.pt, bmarques@med.up.pt

Alberto Freitas\textsuperscript{1,2}

\textsuperscript{1} Department of Biostatistics and Medical Informatics, Faculty of Medicine, University of Porto, Portugal

\textsuperscript{2} CINTESIS - Center for Research in Health Technologies and Information Systems, University of Porto, Portugal

alberto@med.up.pt

Keywords: Outliers Detection. Data Mining. Medical Data.

Abstract:  

**Background:** Patient medical records contain many entries relating to patient conditions, treatments and lab results. Generally involve multiple types of data and produces a large amount of information. These databases can provide important information for clinical decision and to support the management of the hospital. Medical databases have some specificities not often found in others non-medical databases. In this context, outlier detection techniques can be used to detect abnormal patterns in health records (for instance, problems in data quality) and this contributing to better data and better knowledge in the process of decision making.

**Aim:** This systematic review intention to provide a better comprehension about the techniques used to detect outliers in healthcare data, for creates automatisms for those methods in the order to facilitate the access to information with quality in healthcare.

**Methods:** The literature was systematically reviewed to identify articles mentioning outlier detection techniques or anomalies in medical data. Four distinct bibliographic databases were searched: Medline, ISI, IEEE and EBSCO.

**Results:** From 4071 distinct papers selected, 80 were included after applying inclusion and exclusion criteria. According to the medical specialty 32% of the techniques are intended for oncology and 37% of them using patient data. Considering only articles that used administrative medical data, 59% of the techniques were statistical based.

**Conclusion:** The area with outliers detection techniques most widely used in medical administrative data is the statistics, when compared with techniques from data mining such as clustering and nearest neighbor.

1. **BACKGROUND**

Medical databases generally involve many different types of data and produces a lot of information (Kumar et al., 2008). The data typically consists in records which may have several different types of features such as patient age, blood group, weight, clinical images, patient diagnoses, lab test results and other details from patient treatments (Chandola et al., 2009).

Patient medical records include many electronic entries related to patient conditions and treatments and their laboratory results. These are useful in providing a better picture about the individual patient, however, the benefits of such data in decision support or in discovery of new clinical knowledge are far from being exhausted (Hauskrecht et al., 2007).

In latest years, with the exponential development of information technology in hospitals, the volume of medical data has increased significantly. At the same time, new interest in the analysis of this information has emerged, taking place not only as a source for clinical decision making and research in epidemiological studies, but also to support of hospital management (Silva-Costa et al., 2010).
The value assigned to these data is directly related to their quality. This way, as higher is the quality of data, higher is its utility (Arts et al., 2002). For healthcare organizations, this kind of information is essential in providing healthcare, as well as a consistent and careful financial management. On most systems, the quality of data is completely neglected, this way, could become a nightmare, when users of the information produced are not aware about its veracity and quality (Silva-Costa et al., 2010).

In health, this quality is even more important, since a wrong decision by a clinician, may even lead to death of the patient or, in a most extreme case, may also have a more global reach, when this weak quality is associated with the exchange of documents. A professional takes the decision based on surveys that sometimes and without him aware that they are not the same patient that he is addressing. For example, a study on the integration of hospital information systems (Cruz-Correia et al., 2006), it stated that about 0.1% (423 in 391,258) of the documents associated with the process of clinical patients in a hospital central contain identifying information wrong.

Data mining a knowledge discovery in medical databases are not substantially different from mining in other types of databases. There are some particularities in medical databases that are absent in non-medical database (Cios, 2001).

One of these particularities are that the physician’s interpretation of images, signals, or any other clinical data, is written in unstructured free-text, and because of this, standardize is very difficult. Other feature of medical data mining is that the underlying data structures of medicine are poorly characterized mathematically, as compared to many areas of the physical sciences. Physical scientists could substitute data into formulas, equations, and models that reflect the relationships among their data (Cios, 2001).

### 1.1 Outliers Detection in Medicine

Outlier detection is very important to medicine, because this data translate to significant information and often critical data (Chandola et al., 2009). In healthcare databases, outlier detection techniques are used to detect anomalous patterns in patient records which can contain valuable data as, for instance, symptoms of a new disease.

There are many definitions for outliers which differ in words found in different studies (Laurikkala et al., 2000). According to Barnett and Lewis, an outlier can be defined as “an observation, or subsets of observations, appearing to be inconsistent with the remainder of that set of data” (Barnett and Lewis, 1994). In other words, an outlier is an element that deviates from a standard set of data from which it belongs. However, an outlier is always an element of a group. An element is said outlier when compared to a standard, therefore, an element can be called outlier compared to the standard X and not an outlier compared to standard Y (Silva, 2004).

#### 1.1.1 Techniques

The outliers detection techniques debated in this study are:
- **Statistical**: Statistical techniques fit a statistical model, usually for normal behavior, to the given data and then a statistical inference test is applied to determine if an unseen instance belongs to the model or not. Instances that have a low probability to be generated from the learnt model, based on the applied test statistic, are declared as outliers (Chandola et al., 2009).
- **Clustering**: Is used to group similar data into clusters. Even though clustering and outliers detections appear to be fundamentally different from each other, several clustering based outlier detection techniques have been developed.
- **Classification**: This technique is used to learn a model from a set of labeled data instances and, then, classify a test instance into one of the classes using the learned model. Classification based anomaly detection techniques operate in a similar two-phase: the training phase learns a classifier using the available labeled training data and the testing phase classifies a test instance as normal or anomalous using the classifier (Chandola et al., 2009).
- **Nearest Neighbor**: Require a distance or similarity measure defined between two data instances, that can be computed in different ways (Chandola et al., 2009). Techniques based on this approach can be broadly grouped into two categories: techniques that use the distance of a data instance to its nearest neighbor as the anomaly score and techniques that compute the relative density of each data instance to compute its anomaly score.
- **Mixture Models**: Mixture models comprise a finite or infinite number of components, possibly of different distributional types, that can describe different features of data (Marin et al., 2005). In statistics, a mixture model is a probabilistic model for density estimation.
using a mixture distribution. A mixture model can be regarded as a type of unsupervised learning or clustering.

- **Spectral**: Try to find an approximation of the data using a combination of attributes that capture the bulk of variability in the data (Chandola et al., 2009). This technique defines subspaces in which the anomalous instances can be easily identified.

### 1.2 Aim

This systematic review aims to provide a better comprehension about the used techniques to detect outliers in administrative healthcare data, to create automatisms for those methods in order to facilitate the access to information with quality, to make a better decision by managers in health.

### 2. METHODS

#### 2.1 Eligibility Criteria

We defined a pair review, independently, involving the two reviewers in the study. The selection was based on the studies titles and abstracts. The pair review considered eligible articles that contained the following criteria:

- The inclusion covered the topics outlier detection, anomaly detection, extreme data or gross error, and the topics healthcare, health, medical, medicine, clinical or patient.
- The exclusion criteria covered only the articles that described the technique or method used.

Only selection by both reviewers was considered adequate. In case of disagreement the decision was based on a consensus meeting between the reviewers.

To maximize specificity, a preliminary analysis was performed with 20 papers, to evaluate and synchronize the reviewers.

#### 2.2 Review Team

The review team was composed by a Computer Scientist, Juliano Gaspar, and a Medical Doctor, Emanuel Catumbela, advised by a Computer Scientist with expertise in medical informatics, Professor Alberto Freitas.

The statistical analysis was performed by a Statistician, Bernardo Marques, with PASW version 18®.

#### 2.3 Search methods

The search for studies was performed between May and June 2010 in bibliographic databases. Since there were no specific standardized MeSH terms, we developed a search string that includes the concepts of outlier or anomaly detection in healthcare. We decided to not include restrictions on language or date of published articles. Four distinct bibliographic databases were searched: Medline (via Pubmed), ISI (ISI Web of Knowledge), IEEE (IEEE Xplore) and EBSCO (EBSCOnhost® databases). The query search string used in each database was:

```
(((outlier or outliers) and (detect or detection or observation or analysis)) or ((anomalous and data) and (detect or detection or observation or analysis)) or "extreme data" or "gross error" or "anomalous record" or "anomalous register") and (medical or medicine or clinical or patient or care or health)).
```

#### 2.4 Definition of variables

The variables analyzed in this preliminary review are:

- **Outliers detection techniques**: statistical, clustering, classification, nearest neighbor, mixture models and spectral.
- **Data type**: type of data used in the use of each technique (e.g.: images, biosinal, patient data).
- **Medical domain**: which defines the medical specialty or domain of medicine used in the study is employed.
- **Clinical stage**: phase clinical notes that the study is applied to diagnosis, prognosis, treatment, or for assessing outcomes and performance.

Were also collected information if the data used were primary, secondary or simulated, the country where the study was conducted and the year of publication.

### 3. RESULTS

The design of this systematic review showed the following results: the search method found 2697 articles in Medline, 1169 in ISI, 185 in IEEE and 414 in EBSCO, a total of 4465 articles. After eliminating duplicate articles 4071 were selected.

As a result, a total of 177 out of 4071 articles were selected to be read entirely. In this preliminary study we present and analyze 80 articles. Figure 1 is
a flowchart illustrating the different stages of paper selection.

The agreement rate between reviewers during the phase of studies selection was 72%.

![Flowchart](image)

Figure 1: Flowchart illustrating the papers selection.

The table lists all outlier detection techniques considered in this review. In the 80 articles analyzed were identified 112 techniques, which were classified into 6 categories. The “statistical” technique appears with 55.4% of the studies.

<table>
<thead>
<tr>
<th>Type</th>
<th>N</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>62</td>
<td>55.4</td>
</tr>
<tr>
<td>Clustering</td>
<td>17</td>
<td>15.2</td>
</tr>
<tr>
<td>Classification</td>
<td>14</td>
<td>12.5</td>
</tr>
<tr>
<td>Nearest Neighbor</td>
<td>10</td>
<td>8.9</td>
</tr>
<tr>
<td>Mixture Models</td>
<td>6</td>
<td>5.4</td>
</tr>
<tr>
<td>Spectral</td>
<td>3</td>
<td>2.7</td>
</tr>
<tr>
<td>Total</td>
<td>112</td>
<td>100.0</td>
</tr>
</tbody>
</table>

It can be observed in Table 2 listing the 80 articles reviewed, grouped by techniques that each presents. As described earlier, an article may appear more than once, since some studies show more than one technique.

Table 2: List of the articles by techniques

| Statistical               | (Aalen et al., 2004; Ahdesmaki et al., 2005; Ahlers and Figg, 2006; Alameda and Suarez, 2009; Allen et al., 2010; Asare et al., 2009; Bakshi-Raiz et al., 2007; Beguin and Hulliger, 2004; Bickel, 2003; Booth and Lee, 2003; Branden and Verboven, 2009; Breen et al., 2002; Cho et al., 2008; Clutmann and van de Velde, 2000; Cohen et al., 1996; Comanor et al., 2006; Comowick and Warfield, 2009; Cooney et al., 2003; Englesle et al., 2009; Fomenko et al., 2006; Freifeld et al., 2009; Ghosh, 2010; Ghosh and Chinnaiyan, 2009; Glance et al., 2003; Glance et al., 2002; Glance et al., 2007; Gold and Hoffman, 1976; Grotkjaer et al., 2006; Hanauer et al., 2007; Hayes et al., 2007; Hojjatoleslami et al., 1997; Hu, 2008; Hughes et al., 1997; Irgosen and Arenas, 2008; Jackson et al., 2009; Jacobs, 2001; Kaufmann and Huber, 2010; Kazmierczak et al., 2007; Liu and Wu, 2007; Livesey, 2007; MacDonald and Ghosh, 2006; Mahadevan et al., 2004; Meloun et al., 2004; Model et al., 2002; Nielsen and Hansen, 2002; Nielsen et al., 2001; Oh and Gao, 2009; Ohlssen et al., 2007; Penny and Jolliffe, 1999; Penny and Jolliffe, 2001; Read, 1999; Rochelson et al., 2006; Rubin and Chinnaiyan, 2006; Ryan, 2009; Song and Wyrwicz, 2009; Tomlins et al., 2008; Van Leemput et al., 2001; Vankeerberghen et al., 1995; Vellidio and Lisboa, 2006; Whitley and Ball, 2002; Wu, 2007; Zervakis et al., 2009) |
| Clustering             | (Aggarwal and Yu, 2005; Arzmandian et al., 2007; Beguin and Hulliger, 2004; Bickel, 2003; Duan et al., 2009; Freifeld et al., 2009; Gold and Hoffman, 1976; Goovaerts and Jacquez, 2004; Grotkjaer et al., 2006; Hibbs et al., 2005; Irgosen and Arenas, 2008; Jackson et al., 2009; Janeja and Atluri, 2009; Koufakou and Georgiopouls, 2010; Mramor et al., 2007; Vellidio and Lisboa, 2006; Yang et al., 2007) |
| Classification         | (Aggarwal and Yu, 2005; Baker and Jackson, 2008; Cardoso et al., 2007; Comwwick and Warfield, 2009; Gold and Hoffman, 1976; Grotkjaer et al., 2006; Kazmierczak et al., 2007; Law et al., 2001; Lopes et al., 2003; Mahadevan et al., 2004; Oh and Gao, 2009; Ohlssen et al., 2007; Whitley and Ball, 2002; Zervakis et al., 2009) |
| Nearest Neighbor       | (Antao et al., 2008; Beguin and Hulliger, 2004; Chen et al., 2008; Duan et al., 2009; Freifeld et al., 2009; Goovaerts and Jacquez, 2004; Irgosen and Arenas, 2008; Jackson et al., 2009; Janeja and Atluri, 2009; Koufakou and Georgiopouls, 2010) |
| Mixture Models         | (Ghosh, 2010; Ghosh and Chinnaiyan, 2009; Hu, 2008; Lopes et al., 2003; Model et al., 2002; Penny and Jolliffe, 2001) |
| Spectral               | (Cohen Freue et al., 2007; Hubert and Engelen, 2004; Song and Wyrwicz, 2009) |

According to the “clinical stage”, this study showed that 81.3% of the techniques to detect outliers for the diagnostic, while only 3.8% for treatment.

Table 3 lists the data types more frequently found in this study. They are, respectively, “Patient Data” (37.5%) and “Genomic Data Sets” (32.1%) followed by “Images” (15.1%) which includes, for example, the two-dimensional images, three-dimensional, electroencephalography, magnetic resonance imaging and ultrasonography.
This review shows that, considering only administrative medical data, statistical techniques are the most commonly used (59%), followed by clustering (15%) and nearest neighbor (15%). Even considering only administrative medical data, the techniques mainly use data such as "patient data".

Methods used to detect outliers in healthcare databases are mostly applied for diagnosis (81%). Considering medical specialty, oncology and genetic studies are the most applied.

Thus, we conclude that statistical techniques are widely used in administrative medical data, and that techniques from data mining such as clustering and nearest neighbor are still little used in this context. There is a considerable field for developing techniques to detect outliers in medical data management, based on clustering and nearest neighbor, beyond of the statistics techniques already widespread.

### 4.1 Limitations

A major difficulty observed in this systematic review was the variety of names used to define outliers. The lack of objectivity and clarity to describe the methods and techniques used in some articles, demanded the reviewers to have a special attention to distinguish which outlier detection technique was used by the author in his study.

### 4.2 Future Works

In future work, we will implement a prototype for detection of outliers in medical administrative databases using clustering techniques, nearest neighbor techniques, beyond statistical techniques. The aim is to improve the process of decision making the department directors and hospital administrators, providing them with data with more quality and accuracy.

### ACKNOWLEDGEMENTS

The authors would like to thank the support given by the research project HR-QoD - Quality of data (outliers, inconsistencies and errors) in hospital inpatient databases: methods and implications for data modeling, cleansing and analysis (project PTDC/SAU-ESA/75660/2006).
REFERENCES

systems: How it may ensure and improve the quality of data. Stud Health Technol Inform 21, 176-182.


