Physica Medica 60 (2019) 91–99

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Analysis of a CT patient dose database with an unsupervised clustering approach

O. Rampado[a,*], L. Gianusso[a], C.R. Nava[b], R. Ropolo[a]

[a] S.C. Fisica Sanitaria, A.O.U. Città della Salute e della Scienza, Corso Bramante 88/90, 10126 Turin, Italy
[b] Università della Valle D'Aosta, Department of Economics and Political Science, Strada Cappuccini 2A, 11100 Aosta, Italy

ARTICLE INFO

Keywords:
Patient dose archive
Cluster analysis
CT protocol optimization
Outlier analysis

ABSTRACT

Purpose: This study investigated the benefits of implementing a cluster analysis technique to extract relevant information from a computed tomography (CT) dose registry archive.

Methods: A CT patient dose database consisting of about 12,000 examinations and 29,000 single scans collected from three CT systems was interrogated. The database was divided into six subsets according to the equipment and the reference phantoms in the definition of the dose indicators. Hierarchical (single, average, and complete linkage, Ward) and not hierarchical (K-means) clustering methods were implemented using R software. The suitable number of clusters for each CT system was determined by analysing the dendrogram, the within clusters sum of squares, and the cluster content. Summary statistics were produced for each cluster, and the outliers of the dose indicator distribution were investigated.

Results: Ward clustering identified the most common combinations of scanning parameters for each group. The optimal number of clusters for each CT equipment system ranged from 5 to 15. The main diagnostic applications were then extracted from each cluster. Outlier analysis of the dose indicator distribution of each cluster revealed potential improper settings that resulted in increased patient dose.

Conclusions: Clustering methods applied to CT patient dose archives provide a quick and effective overview of the main combinations of currently used exposure parameters and the consequences for dose indicator distributions, also when protocol labels and/or study descriptions are not homogeneous.

1. Introduction

Radiation dose monitoring systems in diagnostic radiology generate large amounts of data, with potential benefits for the practical implementation of radiation protection principles for the optimization and justification of patient exposure [1]. Structured reports containing detailed information on radiation dose allow for deep analysis of the consequences for the exposure parameters of each radiological acquisition on patient dose. In computed tomography (CT), these archives provide an opportunity to interpret the variability of dose indicators, which is often high owing to various causes: multiple scans with different exposure parameters within the same protocol, operation of automatic dose reduction devices, and operator-related changes to the protocols [2]. As a consequence, reviewing and identifying optimization actions is often a rather complex task [3].

Summary statistics of dose indicator distributions are generally calculated by grouping studies and scans according to textual fields, such as the study description, the protocol name, and/or the individual series labels. A description of the study can be generated by the radiological information system or edited by the operator, which often results in different textual values for the same type of study. The lack of standardized nomenclature for the designation of CT studies and protocols makes it difficult to reliably extract dose distributions by simply dividing the data according to protocol name. Furthermore, combined with the individual scans of each protocol, there is a basic set of exposure parameters for which one needs to verify which of them are invariant in all protocol applications.

Several techniques taken from "Big Data" [4] can be used to extract relevant information. In this context, data mining techniques, defined as a process of selection, exploration, and modelling of large masses of data, are employed to discover regularity or relationships not known *a priori*, and to obtain a clear and useful result for the database owner. In patient radiation dose monitoring, the term data mining has only been used with the limited meaning of "data extraction" [5–8]. It is worth

[*] Corresponding author.
E-mail addresses: orampado@cittadellasalute.to.it (O. Rampado), lgianusso@cittadellasalute.to.it (L. Gianusso), c.nava@univda.it (C.R. Nava), rropolo@cittadellasalute.to.it (R. Ropolo).

https://doi.org/10.1016/j.ejmp.2019.03.015
Received 30 November 2018; Received in revised form 19 February 2019; Accepted 17 March 2019
Available online 29 March 2019

noting that these data were already available, although stored in different archives and not structured in one dataset [5]. For example, Ikuta [6] described a data mining technique in nuclear medicine to automatically extract from the exam reports the values of the contrast activity and the radiopharmaceutical. They applied a similar approach to CT examinations [7], by which they extracted the dosimetric indicators from the reports in the form of an image combined with anatomical districts through textual fields describing the exam type and the protocol. Wang [8] used the term "arbitrary data mining" to describe an algorithm that memorizes, monitors, and produces alert messages when predetermined dose indicators are exceeded. A "knowledge based" module is also described to match, based on a correspondence table, the labels of the information fields of interest (dicom tag) of the various manufacturers to those of a standard nomenclature.

Among multivariate and data mining techniques, cluster analysis [9] enables the allocation of observations in subsets, more properly named clusters or groups. Observations should be as similar as possible to each other when they belong to the same group and as different as possible when included in distinguished clusters. The allocation of each observation takes advantage of algorithms, usually without initial constraints on the final number of clusters required (hierarchical approach), based on similarity or dissimilarity measures between observations, namely, suitable metrics (or distances). Differently from commonly used human classification, this approach allows for unsupervised self-learning directly on available observations and not grounded on examples or textual descriptive fields or predefined variable intervals of values. Based on the distance computed over the multiple variables for each individual, cluster analysis can identify previously unknown homogeneous groups.

Cluster analysis has widespread application. For example, targeted marketing focuses on specific audiences identified by customer segmentation. Other areas where cluster analysis is exploited include image pattern recognition, computer security systems, and analysis of biological samples.

The aim of the present study was to evaluate the utility and the effectiveness of cluster analysis to extract relevant information from a CT patient dose database. The extracted information could provide an overview of the dosimetric performance of a CT scanner and of eventual situations requiring attention for potential improvement. In addition, we wanted to determine whether this approach, the first of its kind to our knowledge, can quickly highlight the combinations of exposure parameters commonly used in clinical practice and the relative distribution of dose indicators or deviations from standard protocols.

2. Material and methods

2.1. Dataset characteristics

The dataset was generated using Dosewatch dose recording software (GE Healthcare, Milwaukee, WI, USA) [10] to collect the information of each CT scan performed over the course of 1 year (2017) on three equipment systems (Table 1). For each irradiation event, 101 variables were recorded in the archive. The variables refer to: site and equipment data, patient data, study code and time, descriptive fields of the exam type and acquisition protocol, technical parameters of exposure, and dose indicators. The values of these fields were taken automatically from the Dicom Header of the images and from the Radiation Dose Structured Report using the dose recording software. Six data subsets were identified for the data collected by the same equipment and with the same reference phantom in the definition of the dose indicators (head or body). Table 2 presents the database fields considered in the cluster analysis. The dataset was exported as a csv file from the Dosewatch software and imported in RStudio, an integrated development environment for R software (www.rstudio.com, version 1.1.419), to perform statistical and multivariate analysis. Cluster analysis was performed using the dedicated functions "hclust" and "kmeans" in the base package "stats" (version 3.5.2).

2.2. Basic steps in cluster analysis

Each record in the database represented a multivariate variable of collected exposure parameters (Table 2). The parameters in bold were selected for cluster analysis, given their importance and impact on patient dose. The analysis excluded exposure parameters susceptible to significant variations due to individual patient anthropometric characteristics, such as the average anodic current, and often defined by the automatic modulation system (*mAA*). Similarly, table height and absolute position at the beginning and the end of the scan cannot be considered as being specific to the single scan mode since they depend on good technique and on the work of the single operator.

In order to perform cluster analysis on patients with quantitative variables, a suitable metric should be selected to measure the similarity shared by two observations and to initialize cluster analysis. Euclidean distance is most commonly used, as in this case. Let X be the data matrix of n patient scans measured across k variables, where X_i denotes the k-dimensional vector of the i-*th* observation and x_{ih} is its generic element with $h = 1, \cdots, k$. The Euclidean distance measure between the i-*th* and j-*th* observations is:

$$d_{ij} = \left[\sum_{h=1}^{k} (x_{ih} - x_{jh})^2\right]^{\frac{1}{2}} \tag{1}$$

Euclidean distance is sensitive to different variable measurement scales. This could affect cluster analysis by introducing bias in group formation. In other words, single variables with high numerical values and absolute proportional variations may drive the clustering. In this case, we can observe different variation ranges for the tube current variable *mAM* and the pitch factor *Pit.* To overcome this issue, the variables were standardized by linear transformation (i.e., subtract the sample mean and divide by the sample standard deviation to obtain variables with null sample average and unitary sample variance) to ensure that they contribute in the same manner to cluster analysis.

Hence, cluster analysis can be implemented on the constructed distance matrix following a hierarchical or not hierarchical approach [9]. Among the not-hierarchical methods, we selected the so-called K-means or method of centroids in which the resulting clusters can be uniquely associated with individual virtual observations calculated as cluster centroids. Typically, the centroid is the average of variables computed only for observations composing the cluster. The distance between the i-*th* observation x_i and the h-*th* cluster C_h is reduced to the Euclidean distance between x_i and the cluster centroid c_h, $d(x_i, c_h)$. The algorithm steps are iteratively implemented to minimize the distance of the individual observations from the centroid of a cluster and to maximize the distance from the centroids of the other clusters. The number

Table 1
CT equipment considered in this study, related number of examinations and of single series.

CT equipment	Number of slices	Maximum z axis collimation	Number of examinations	Number of single series
GE Brightspeed Elite	8	20 mm	3748	8737
GE Optima 660	64	40 mm	4990	10,246
GE Revolution CT	256	160 mm	3570	9983

Table 2
Single information fields archived in the CT patient dose database. Variables used in the cluster analysis are denoted in bold.

Information field category	Single information fields
Site and equipment data	Site *(Sit)*, Institution name *(Inst)*, Manufacturer *(Man)*, Model *(Mod)*
Study code and time	Study date *(Std)*, Study time *(Stt)*, Accession number (Acn)
Description of the study and of the acquisition protocol	Local study description *(Lsd)*, Study Protocol Name *(Stn)*, Series description *(Sds)*, Series protocol name *(Sen)*, Series type ***(Set)***
Patient data	Patient birthdate *(Pbd)*, Age class *(Acl)*, Patient sex *(Pse)*, Patient weight *(Pwe)*, Patient size *(Psi)*, BMI *(BMI)*, SSDE LAT *(Slt)*, SSDE AP *(Sap)*, SSDE effective diameter *(Sed)*
Exposure parameters	Total number of irrad. Event (*Tni*), Num. Series Non SmartPrep or Localizer *(Nsr),* Start slice location *(Ssl),* End slice location *(Esl)*, Scanning length ***(Scl)***, Gantry detector tilt ***(Gdt)***, Table height (*The*), Tube voltage ***(kVp)***, Max X-ray tube current (mA) ***(mAM)***, average X-ray tube *current (mAA),* Exposure time per rotation ***(Etr)***, Ten time GE noise index ***(tNI)***, Pitch factor ***(Pit)***, Slice thickness ***(Sth)***, Spacing between slices *(Sps)*, Nominal single coll. width ***(scW)***, Nominal total coll. width ***(tcW)***, Table speed ***(Tsp)***, Table feed per rotation (*Tfr*), Iterative recon level ***(Irl)***, Convolution kernel *(Cke)*
Dose indicators	Mean CTDI_{vol} *(CTDI),* CTDI phantom type *(Pha),* DLP *(DLP)*, Total DLP *(tDLP)*, SSDE *(SSDE)*

of clusters is prior information; the method is used when dealing with huge amounts of observations, given that it is very fast, with a calculation time that varies from a few seconds to a few minutes, depending on dataset dimension and complexity [11].

Differently, in hierarchical clustering methods, the aggregation process starts from the original dataset and considers each observation as a singleton. Groups are then progressively constructed based on the computed distance matrix of observations and clusters generated in the previous step. The process terminates when all available observations belong to the same group. The appropriate number of clusters can be determined *a posteriori* by evaluating, for instance, the shape of the dendrogram (i.e., the graphical representation of the aggregation resulting from clustering). A specific method should be selected to compute at each step t ($t = 1, \cdots, n - 1$) the reduced distance matrix and the associated clustering. If Euclidean distance is the metric used in the analysis, it is not the only criterion with respect to which the distance between clusters and singletons or other clusters can be defined. The choice of a specific method might result in different partitioning of the same sample [12].

Moreover, there are also other approaches in empirical analysis. The single linkage or the nearest neighbour method computes the distance between two groups as the smallest distance between elements that compose them. The average bond method uses the average distances between all observations belonging within the two clusters. The complete linkage method selects the maximum distance between units making up the clusters. Finally, like the K-means approach, the Ward method uses an objective function to create clusters that present maximum internal cohesion and maximum external separation [13]. The total deviation T computed over the k variables is considered. It corresponds to n times the trace of the data matrix, decomposed into its two components: the deviance within clusters W and the deviance between groups B. At each iteration, the formation of a new cluster is driven by the simultaneous minimization of the deviance within clusters W and the maximization of the deviation between clusters B. In other words, it preserves maximum homogeneity within and maximum heterogeneity between clusters. In hierarchical methods, the optimal number of clusters that will be considered is not known *a priori* but rather depends on the method selected.

2.3. Clustering method and selection of number of clusters

The clustering methods were applied to the extracted data to group homogeneous scanning modes and to identify the approach most efficient in achieving high compactness within and high separation between clusters. In general, the number of clusters sought varied initially between 2 and 30 for each method and the total deviance inside them was evaluated (within sum of squares). We evaluated the dendrogram, i.e., the graphical representation of classification trees that describe the agglomerative processes of hierarchical clustering. The tree branches represent successive groupings of the statistical units. Distances are shown on the ordinate axis, illustrating the distances to which the different statistical units are joined, also called “aggregation indexes”. Based on the total deviance within the clusters and the dendrogram, the optimal number of clusters for each subgroup was decided.

2.4. Analysis of cluster content

The method was selected, the number of clusters was identified, and the main distinct partition features were analyzed. Graphical and tabular format outputs were exploited using a routine written in R. Each cluster was initially named with a progressive integer. Information was derived for each cluster and stored in different csv files (of tiff for the plot case) containing:

- scans and all associated observed variables (also those not included in the cluster analysis) of the subset matched with the cluster id the observation belonged to;
- summary statistics split by clusters of disposable variables. In detail, the relative frequency of each modality in the presence of categorical variables, while minimum, maximum, first and third quartiles, mean and median in the presence of quantitative variables were stored;
- summary statistics for exposure parameters and dose indicators within clusters;
- contingency tables of the id cluster against the series protocol name, the series description, and the study description. These were particularly useful for associating the main diagnostic applications to each cluster and to each exposure parameter combination;
- the box and whiskers plots of exposure parameters used in the cluster analysis, of the average anodic current and of the three dose indicators ([CTDI] computed tomography dose index, [SSDE] size-specific dose estimate, and [DLP] dose-length product);
- the list of outliers for the distribution of CTDI dose indicators.

The generation of these outputs required an R runtime of less than 1 min for each subset, including the cluster analysis.

3. Results

3.1. Clustering method and number of clusters selected

Fig. 1 shows the dendrograms of the four hierarchical methods illustrated in the previous section based on the (scaled) subset relative to the GE Optima 660 tomograph and head anatomical district. Relative distances between partitions at each iteration are reported on the vertical axis. The range of values was the same for the single, medium, and complete link, while the numerical scale was completely different for the Ward method, as was the shape of the dendrogram. Partitions were progressively merged into a unique cluster within a very narrow range of relative distances for the single method. As expected, the

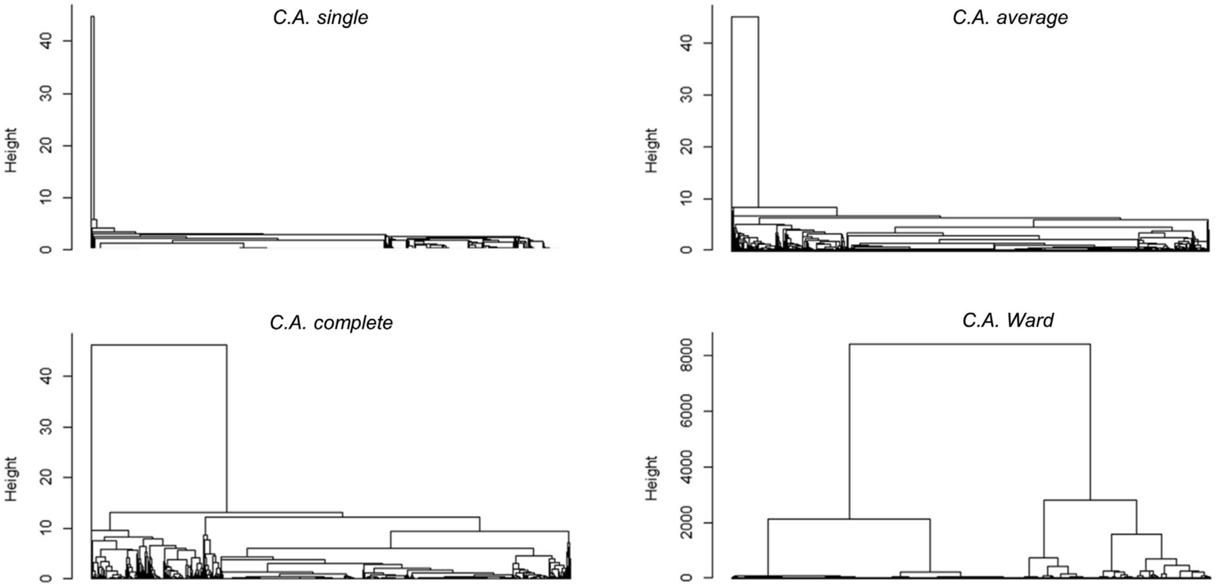

Fig. 1. Dendrograms of the subset head of the GE Optima 660 CT obtained with four different hierarchical clustering methods.

dissimilarities were progressively greater for the average, complete, and Ward methods [12]. The literature often recommends selecting the number of clusters that "cuts" the dendrogram at the longer "branches", i.e., the greater relative distance variations corresponding to the union of smaller dimensional clusters. A choice of only two clusters would, therefore, appear natural. However, for this specific application we wanted to highlight very homogeneous clusters and evaluate the impact of minor differences on the dosimetric indicators between only some of the component variables. In the Ward dendrogram in Fig. 1, one could then orientate towards an extraction of seven clusters.

The trends of total internal deviance within the clusters and the number of clusters varied for the four methods and for the non-hierarchical K-means method (Fig. 2a). Common values can be seen between the Ward method and the K-means method for some numbers of clusters and for numbers greater than 25. The decreasing trend is a common result in cluster analysis. A criterion normally used to choose the number of clusters to extract is the so-called "Elbow" method [14]. The Elbow method examines the percentage of deviance explained as a function of the number of clusters: one should choose a number of clusters so that the addition of another cluster does not improve modeling of the data. As shown in Fig. 2, the first clusters add a lot of information (i.e., they explain a lot of deviance), with a sharp decrease in total deviance within the clusters, but marginal gain decreases at some points, resulting in an angle (elbow) in the graph. The number of clusters should be chosen at this point, hence the "Elbow criterion", although it is often not so clear how to exactly position this threshold. For this study, we considered a number of clusters that bring the total internal deviance to the clusters below 20% of the total deviance.

For the subsets in Fig. 2 the choice fell to seven clusters with the Ward method for the GE Optima 660 CT and to 15 clusters for the GE Revolution CT. The subsequent validation carried out with analysis of the actual content confirmed the goodness of this choice.

Analysis of the trend of the total internal deviance within clusters was carried out for all six subsets based on the three tomographs and the two reference districts. Differences across subsets are clearly present due to the nature of the data, technological characteristics, and clinical applications of the equipment. However, highlights for the example in Fig. 2 remain valid also for the remaining subsets, with the number of selected clusters between five and 15 (Table 7). The Ward method was found to be the one that ensured the smallest total internal deviance with the smallest number of clusters.

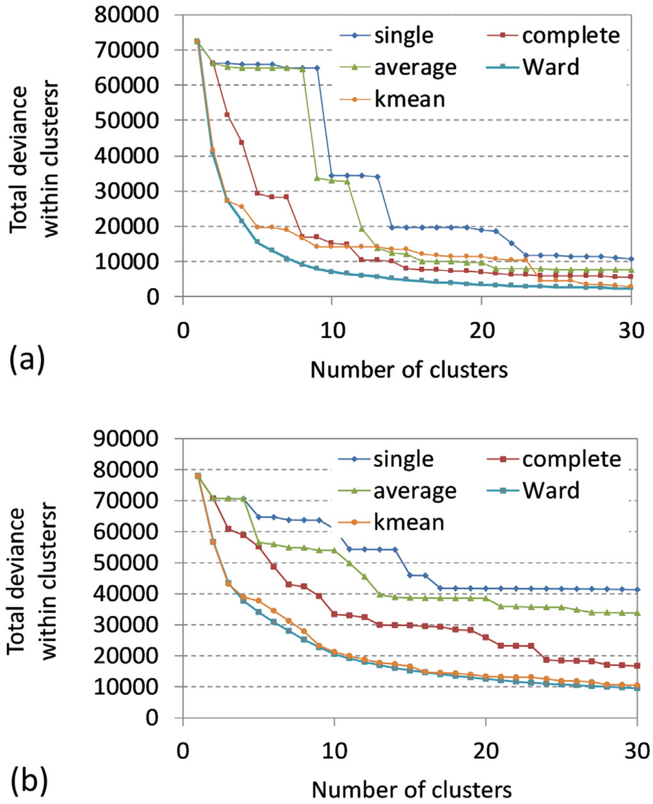

Fig. 2. Total deviance inside clusters vs. number of clusters for the subset head of the GE Optima 660 CT (a) and the subset body of the GE Revolution CT (b) obtained with four different hierarchical clustering methods and the K-means method.

3.2. Cluster naming and relative dose distribution

To assign the current prevalent uses to the clusters of scans, we evaluated the distribution of textual descriptive fields with respect to the identified clusters. We examined the representative example of the head scans for the GE Brightspeed Elite TC. Table 3 presents the combined frequencies between cluster id and the main protocols. The facial bone (2.1) and the upper limb (4.7) protocols principally characterized

Table 3
Contingency table with the number of scans having the indicated series protocol names and belonging to the different clusters extracted, for the CT GE Brightspeed and for the subset of head scans.

Protocol name	Cluster				
	1	2	3	4	5
1.3 Head Trauma – C1 C2	2	612	622	636	0
1.1 Standard Head	0	0	449	458	0
2.1 Facial bone	97	1	0	0	0
4.8 Upper limbs (small joint)	0	80	0	0	0
4.4 Small joint HR	2	70	0	0	0
4.7 Upper limbs	63	1	0	0	0
1.2 Helical head	0	0	0	0	55
3.1 Cervical spine	3	52	0	0	0

Cluster 1. Although they are two different anatomical districts, clustering tended to merge them into a single cluster since the scanning parameters were similar. Cluster 5 corresponded to the helical head (1.2) protocol.

Differently, protocol names combined with scan clusters exhibited large overlaps for the first two protocols (1.1 and 1.3). This was particularly evident in Clusters 3 and 4, where protocols 1.3 and 1.1 are similar for the two clusters. In order to combine these clusters with current practice, we found it useful to report highly frequent exposure parameter values for each cluster and to trace them to the known aspects of the radiological technique used. For instance, in Cluster 3 a slice thickness of 5 mm is used and there is a median scan length of 9 cm, while in Cluster 4 the slice thickness is thinner (2.5 mm) and the median length is 6 cm. Observing the other reported parameters, we may deduce that in most cases two scan sections are performed sequentially to acquire brain and the posterior fossa images. These two scans are present in the standard skull (1.1) protocol (consisting only of them) and in the C1 C2 trauma skull (1.3), together with the second cluster. The use of the latter is evident when we look at the series descriptions (not reported here): the indications “C1-C2”, “cervical” and “cervical spiral” refer to the scan with spiral acquisition of the cervical spine. Cluster 2 contains also small joint scans, with similar exposure parameters.

We have intersecting information with a basic knowledge of the commonly used radiological scanning techniques. In some cases, it was useful to inspect the details of the single exam, looking at additional information through the Dosewatch user interface. Following this approach we were able to assign a “name” to the clusters identified for the tomographs, i.e., we were able to identify the areas of prevalent clinical use and combine them with the median exposure parameters (Table 4).

The box and whiskers plots of the individual exposure parameters and the distribution of the associated dose indicators were analyzed to determine whether the group was able to explain the variability of these quantities, which is crucial for radioprotection of the patient. Fig. 3 shows the diagrams for the CTDI and DLP indicators for the five clusters. The homogeneity of the exposure parameter combinations inside each cluster is reflected in the rather compact distribution of the dose indicators. This gives a concise picture of their variability and dose levels associated with the diagnostic applications commonly used. In most cases, the scans of the facial bone, the small joints, and the cervical spine had low local dose (CTDI < 30 mGy) and dose-length product values (DLP < 500 mGy cm). The CTDI was around 50 mGy for the brain scans and around 70 mGy for the posterior fossa, with similar DLP values due to the different scanning length. The highest values for both CTDI and DLP were found for the helical head scan. A large number of outliers were visible and subsequently analyzed.

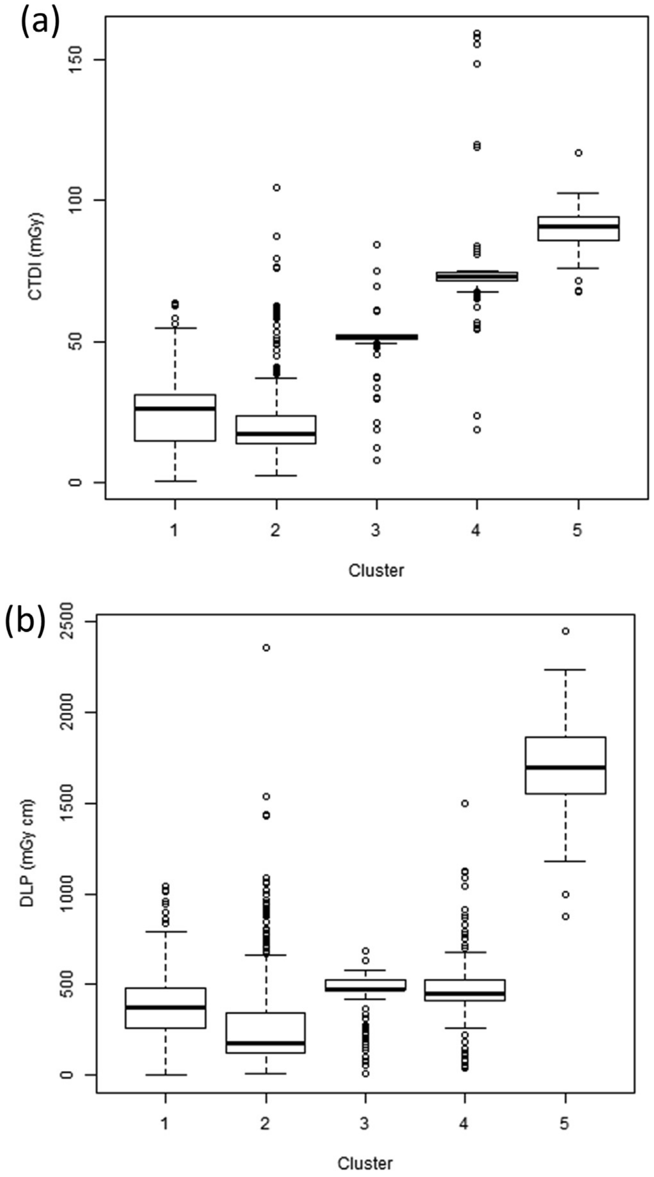

Fig. 3. Box and whiskers plots of the CTDI$_{vol}$ (a) and DLP (b) distributions for the scans of the five clusters extracted from the CT Brightspeed head subset.

We then wanted to know what type of dose indicator distributions would be obtained when considering the protocol name categories, the

Table 4
Number of scans, median exposure parameters and prevalent clinical use for each cluster for the CT GE Brightspeed and for the subset of head scans. The exposure parameters associated to the variable names indicated are defined in Table 2.

Cl	n	Set	kVp	mAM	mAA	Etr	tNI	Sth	Gdt	Pit	tcW	Scl	Prevalent use
1	217	Sp.	100	160	160	0.8	0	1.25	0	0.63	10	157	Facial bone, upper limbs
2	898	Sp.	100	200	130	0.6	140	1.25	0	0.75	5	94	Cervical spine, small joints
3	1097	As.	120	140	140	2	0	5	10.5	0	10	92	Brain
4	1103	As.	120	160	160	2	0	2.5	13	0	5	62	Posteriour fossa
5	55	Sp.	120	382	345	0.8	28	2.5	0	0.63	10	186	Helical head

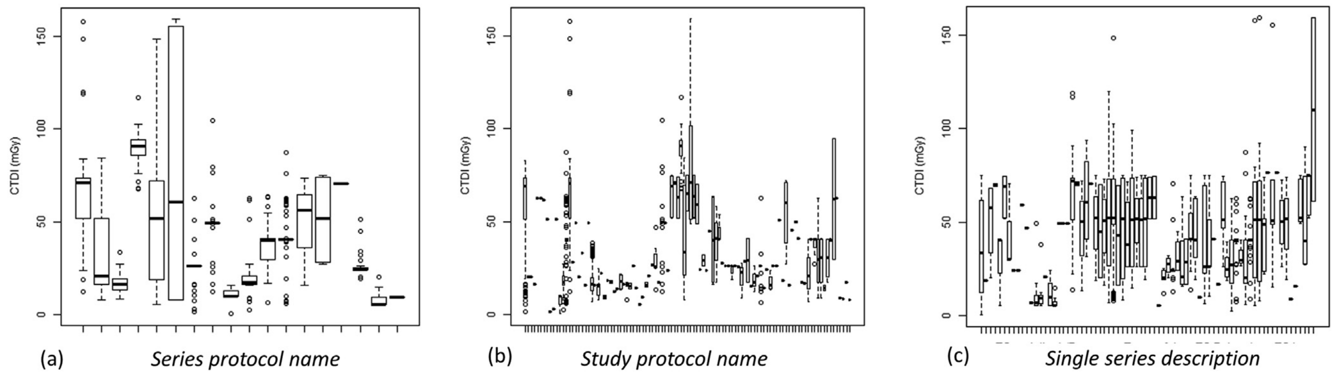

Fig. 4. Box and whiskers plots of the CTDI_{vol} distributions for the different series protocol names, study, and single series descriptions of the CT Brighspeed head subset.

series description, and the study description (or the name of the exam). Fig. 4 shows the relative box and whiskers diagrams: the high number of categories and the frequently dispersed distributions of the indicators make it difficult if not impractical to extract an effective synthetic framework, such as that obtained by cluster analysis, differently from the approach proposed here (Fig. 3).

A second example of the approach comes from GE Revolution CT equipment and the body district. This equipment is a latest-generation tomograph with multiple innovative and improved elements as compared with the other two CT systems. It simultaneously acquires 256 slices on a volume of 16 cm in length with a single rotation of the X-ray tube. This is a particularly important feature for performing cardiac CT. Given the complexity and the number of scans in this subset, and the total deviance within the cluster trend, the partition was constructed considering 15 clusters (Table 5 and Fig. 5).

The first two clusters refer to heart scans: the first cluster contains an acquisition for producing images (some of which also three-dimensional) for planning aortic valve implantation; and the second cluster refers to the diagnosis of coronary arteries and cardiac chambers. The dispersion of relative CTDI values is largely a consequence of the duration of acquisition, which depends on the patient's heart rate. Another major component of examinations carried out on this device concerns cancer patients. During staging, it is often necessary to acquire multiple scans of the same districts along different timelines following injection of a contrast agent: baseline scan, then early arterial (15–20 s after the beginning of contrast injection), late arterial (30–35 s after contrast injection), and portal venous phase (55–60 s after contrast injection). For this type of examination, all the phases are included within the same protocol, theoretically differentiable in the archive only by the series description, which is usually not very specific, however.

Contingency table analysis and evaluation of individual cases allowed us to identify, among prevalent uses, also the different phases of these exams associated with the related clusters. In some cases, the uses could be superimposed on different cluster pairs, such as Clusters 6 and 15 which exhibit two different median parameters. Apart from the first two clusters related to cardiac scans, all the others had median values and most of the third quartiles were < 10 mGy, which is a relatively low dose value given that the latest Italian dose reference levels for chest and abdomen scans are between 15 and 18 mGy [15]. Higher DLP values can be observed for Cluster 12, coherent with lower limb angiography, associated with a median length of 127 cm. Clusters 13 and 14 represent scans of the entire abdomen or the abdominal chest and had greater median scanning lengths than the other clusters.

Fig. 6 presents the box and whisker plots of individual exposure parameters for the 15 clusters in this subset and clearly indicates which parameters are actually used within the individual clusters. In many cases, we observed substantial stability, with constant values of kV, pitch, collimation, and slice thickness. Differently, the NI modulation indices [16] and the percentage values of the ASIR-V iterative algorithm [17] showed more operator-related variations in individual cases.

3.3. Outlier analysis

In the dose indicator distribution of scans with identified clusters, numerous outliers are visualized in the box and whiskers plots (Figs. 4 and 5). Outliers were the single observations with a CTDI_{vol} more than 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile. They might be associated with unjustified high dose values and due to parameter variations made by single operators or to

Table 5
Number of scans, median exposure parameters and prevalent clinical use for each cluster for the CT GE Revolution CT and for the subset of body scans. The exposure parameters associated to the variable names indicated are defined in Table 2.

Cl	n	Set	kVp	mAM	mAA	Etr	tNI	Sth	Irl	Pit	tcW	Scl	Prevalent use
1	122	As	110	440	305	0.28	0	0.63	30	1	160	260	Cardiac CT valvular implant
2	225	As	100	400	299	0.28	0	0.63	50	1	160	160	Cardiac CT
3	884	Sp	120	288	231	0.35	200	1.25	60	0.98	40	351	Abdomen CT Arterial
4	374	Sp	100	431	299	0.28	226	0.63	60	0.99	80	375	Pulmonary embolism
5	201	Sp	120	334	213	0.5	220	1.25	40	0.98	40	309	Abdomen CT four phases portal
6	201	Sp	120	400	269	0.28	230	0.63	50	1.38	40	374	Expiration chest CT
7	702	Sp	120	340	257	0.35	220	1.25	60	0.99	80	358	Angiographic abdomen CT
8	581	Sp	100	328	238	0.5	119	5	40	0.98	40	295	Abdomen CT four phases without contrast or late arterial phase
9	773	Sp	120	304	212	0.5	120	5	40	0.98	40	305	Abdomen CT four phases portal
10	294	Sp	120	148	120	0.5	190	1.25	60	0.52	40	336	Liver CT three phases
11	196	Sp	120	382	162	1	118	2.5	40	0.98	40	244	Neck CT
12	64	Sp	100	404	213	0.5	110	5	40	0.98	40	1271	Lower limbs CT
13	704	Sp	120	456	261	0.5	110	5	40	0.98	40	520	Abdomen CT four phases arterial
14	1322	Sp	120	458	288	0.35	200	1.25	60	0.98	40	665	Chest abdomen pelvis CT portal phase
15	442	Sp	120	433	277	0.35	230	0.63	50	0.98	40	310	HR Expiration Chest CT

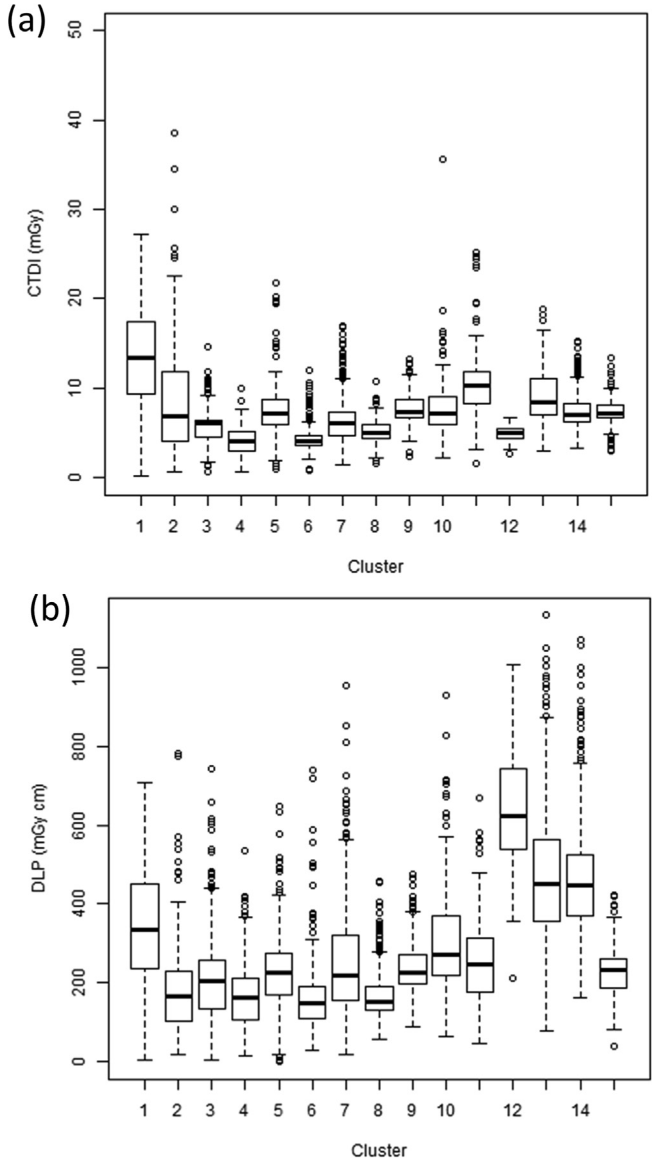

Fig. 5. Box and whiskers plots of the CTDI_{vol} (a) and DLP (b) distributions for the scans of the 15 clusters extracted from the CT GE Revolution CT body subset.

unusual behaviour of the automatic processes in the equipment. It is known, for instance, that a patient positioning error inside the tomograph can lead to incorrect evaluation by the anodic current regulation software [18].

When analysing outliers, we assume that their causes can be grouped into three categories:

1. scans for purposes other than those described as prevalent: it is interesting to check whether they are applications in other clusters with modified parameters or are particular uses that have not given rise to a limited number of autonomous clusters (“different uses”);
2. scans with parameters of the volume to be acquired or in non-standard clinical cases (“special patients”) modified to meet the diagnostic needs of patients wearing prostheses or medical devices;
3. scans with variations in the parameters made by single operators to improve the image quality or consequent to anomalous equipment operation (“modified parameters”).

In order to classify outliers (in terms of CTDI) according to these categories, we analysed them including available information such as exposure parameters, study type, and acquisition protocol (Table 2). Furthermore, to identify those belonging to the first category (scans performed for uses other than those considered as prevalent for the relative cluster), the textual variables for the definition of the prevalent uses were often exploited, then the study description, the protocol name, and the series description. To distinguish between the second and the third category, we analyzed the relative frequencies of variation in the combination of parameter recorded from the images of a sample of scans in Dosewatch software. Useful information about these patients was gathered from interviews with radiologists who use the equipment.

Table 6 shows the frequencies of the outliers for the head subset of the GE Brightspeed tomograph, which had the highest percentage of anomalous values (Table 7). Together with the total observations “n” in each cluster, Table 6 presents “low” outliers – those with CTDI values in the lower part of the distribution – and “high” outliers – observations in the upper part of the CTDIvol distribution. The three categories assumed as determinants of the outliers are shown in the last three columns. A total of 233 outliers were observed (6.9% of the total observations): 67% (157) were high outliers and the remaining 33% (57) were low outliers. The first cluster had six outliers, all were upper limb scans with operator-modified parameters. The second cluster had the largest number of outliers (55%): among the two prevalent uses, 48 outliers were related to the cervical spine and 80 to small joints, mainly of the wrist and the hand. For most uses related to the cervical spine, we observed questionable changes to the basic settings made by the operators (increased kV, reduced modulation index, etc.). We observed maximum levels of beam intensity within the expected range for small joint scans: 26 outliers were present in the third cluster, 9 of which with low values and 17 with high values. The very compact distribution of the remaining observations was a consequence of the use of numerous fixed parameters. The interquartile distance was 2% of the median value. There were only 13 outliers with differences above 20% of the median: 3 had high values, 2 of which were associated with polytraumatized patients in which the parameters were altered to improve image quality due to the presence of devices in the field scan. Similar considerations apply to the fourth cluster: the posterior fossa scans. There were 6 scans with very high CTDI values, about twice the median, resulting from an operator-related increase in the anodic current value. Although there are polytraumatized patients who often require high diagnostic accuracy under critical acquisition conditions, an increase of this magnitude appears doubtful and deserves further study to substantiate its use in cases of actual need.

4. Discussion

The installation and implementation of dose management software allows for useful statistical analysis based on the distribution of dosimetric indicators associated with the names of procedures and protocols. A crucial condition is proper use of the descriptive fields, which have to be uniquely associated with practices for same purposes and methods. This often does not occur, however. When the study description consists of an alphanumeric field completed during the examination, it will provide a clear indication with a text that may present infinite variations, also with respect to possible abbreviations of specific terms. It may happen, for example, that medical operators and technicians use different acquisition protocols with different minimum parameters, so that multiple protocols are often present for the same procedure on the same equipment. In many cases the number of acquisitions carried out or the type of procedure is defined during the course of the exam, following compilation of the fields that will be reported in the recording of the exam and the relative dose indicator.

Given these operational elements, greater homogeneity of the data collected and, in particular, of the database fields for identifying the practices can be achieved in two ways. First all operators should be involved in defining a standardized nomenclature appropriate for the activity and the variations that can occur in work flow. Second,

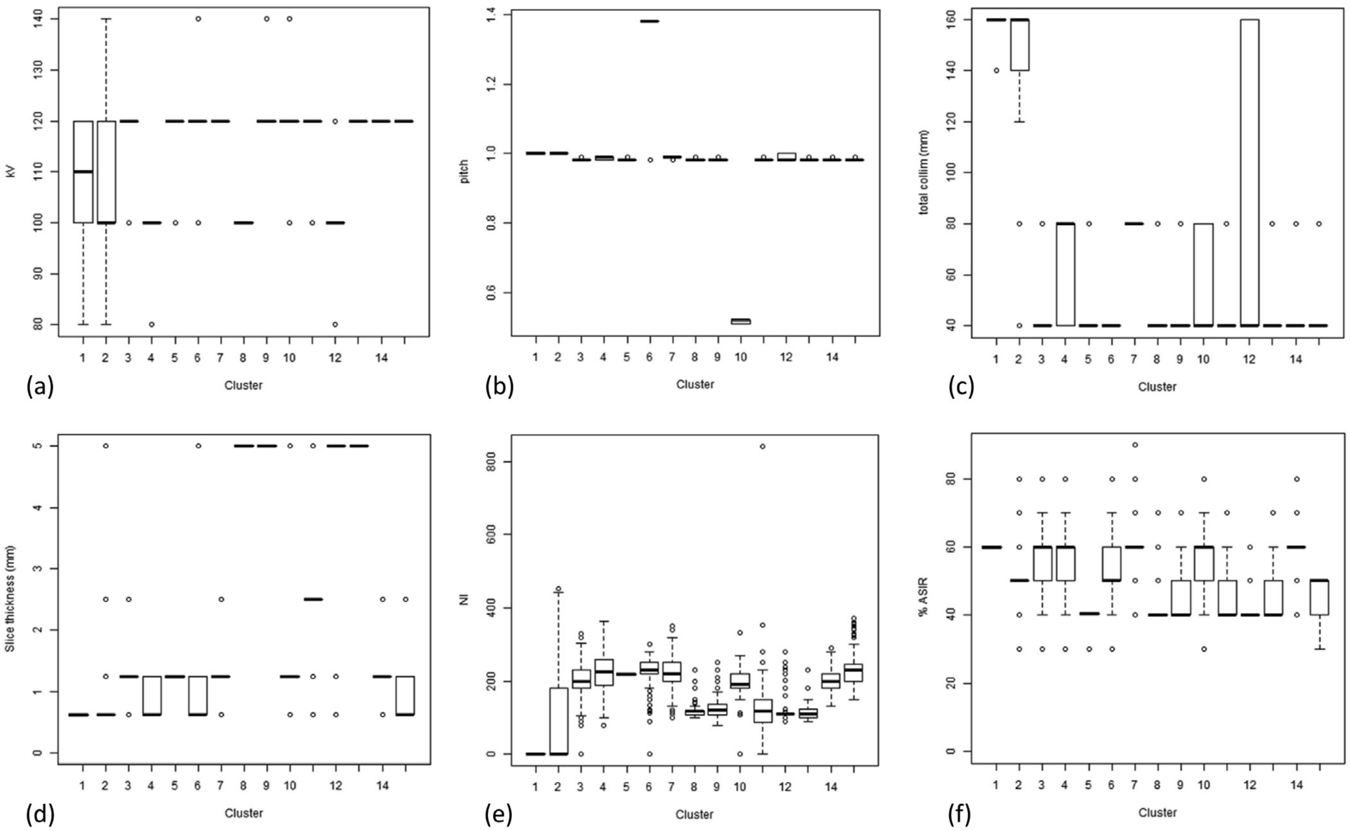

Fig. 6. Box and whiskers plots of exposure parameters for the scans of the 15 clusters extracted from the CT GE Revolution CT body subset. (a) Kilovoltage (b) pitch (c) total collimation (d) slice thickness (e) ten time noise index (f) percentage of iterative level ASIR-V.

Table 6
Frequencies of the different outliers for the subset head of the GE Brightspeed tomograph.

C	Prevalent use	n	n low outliers	n high outliers	n different uses	n special patients	n modified parameters
1	Facial bone, upper limbs	217	0	6	0	0	6
2	Cervical spine, small joints	898	0	128	0	0	128
3	Brain	1097	21	5	0	2	24
4	Posteriour fossa	1103	54	15	0	3	66
5	Helical head	55	1	3	0	0	3

Table 7
Number of clusters and percentages of outliers for the considered subsets.

CT scanners	Subset	Number of scans	Number of clusters	% outliers
GE Brightspeed	Head	3370	5	6.9%
	Body	4812	10	6.4%
GE Optima 660	Head	6019	7	2.8%
	Body	3327	11	1.2%
GE Revolution CT	Head	689	7	0.7%
	Body	7089	15	2.6%

different dose-recording software programs provide tools for grouping records with similar descriptive fields.

With this study, we wanted to determine the effectiveness and usefulness of an alternative approach based on data mining techniques. After an exploratory phase in which we evaluated the applicability of existing methods, we then focused the study on cluster analysis and the tools derived from its implementation. Specifically, we asked whether this tool can actually provide an answer to the following question: what are the combinations of exposure parameters most commonly used in the clinical practice with each tomograph and the resulting distributions of dose indicators? The objectives of this approach are to avoid errors in association with the protocols described above and to highlight the deviations from the basic scanning protocols stored on the equipment.

Another interesting application is the possibility to compare different dosimetric indicators resulting from individual scans in multiphase examinations, such as for different device scans with and without contrast media, which are difficult to extract from the dataset due to the absence of adequate descriptive fields. Finally, we evaluated the effectiveness of cluster analysis to highlight dataset outliers with respect to single observed variables, which may sometimes represent significantly higher patient dose values than the average, in this case as a consequence of improper setting of the scan parameters by single operators. A recent study investigated via cause and effect analysis dose outliers in CT patient dose data [19]. Differently from the approach proposed here, the study was focused only on a single protocol, with a decidedly simpler context than the one explored in our study.

Our results show that the K-means and Ward methods were the most suitable for analyzing the structure of the observed variables. The possibility to synthesize salient information from thousands of tests, including dozens of types of diagnostic studies and different scanning protocols, was achieved with a maximum of 15 clusters and a high homogeneity rate of the 12 characterizing variables and of the radiation dose indicators. An effective field of prevalent clinical use was identified for each device and for each cluster.

Comparison of the distribution of the quantities of interest and analysis of the anomalous values for the different clusters highlighted several possible actions to improve current practices. The time of implementation of the analysis in relation to the amount of information obtained was fully compatible with routine use of this instrument for optimizing patient radiation protection.

Patient height and weight data were often absent in the database, as collected with manual entry by the operators that was not performed constantly. The patient size could be useful for a second level analysis: once the clustering was performed considering the exposure parameters not dependent on patient size and on operation of the automatic exposure adjustment, it could be possible to extract clusters of patients with homogeneous dimensions and to verify the corresponding distribution of dose indicators.

Other possible applications include, for example, analysis of the frequency of the use of technological features during the acquisition of new equipment to appropriately evaluate those characteristics that are really useful in diagnostic imaging. The classification we obtained with cluster analysis lays the foundation for rationalization of acquisition protocols recorded by the equipment control software and the nomenclature of the types of studies carried out, which is also useful for reporting the examination activity.

The recent introduction of standardized nomenclatures, such as Radlex, can improve the consistency of dosimetric registers and facilitate comparison across equipment systems and medical centers. However, the hierarchical structure of such approaches determines in any case a large number of possible radiological procedures. The current Radlex notebook contains about 80 main descriptions of CT imaging procedures, for each of which different attributes are possible, such as the anatomical focus and the reason for the examination, which inevitably lead to the definition of several hundred possible CT studies. The scanning parameters and the resulting dosimetric indicators can be homogeneous even for very different CT studies. This type of information can be obtained with a cluster analysis like the one proposed here. The use of standardized nomenclature and cluster analysis can provide two complementary tools to relate national dose registries and optimize processes more efficiently.

Future areas of focus may be the use of mixed qualitative and quantitative variables, in which descriptive variables hitherto used only for the validation and for the definition of the prevalent uses of the different clusters are integrated into the grouping process. The Euclidean metric of numerical distances is replaced by dissimilarity indices, such as the Gower index [9].

Recent developments in cluster analysis [20] consider a two-level clustering process: the traditional one of observations as used so far and then clustering of variables on the basis of their values in different observations and reciprocal correlations. It can be said that this type of analysis combines analysis of the main components with the clustering process and provides the possibility to extract groups of homogeneous observations and, at the same time, reduces the number of variables by grouping them according to their effective influence on data variability.

Another possible extension of the present study could be Fuzzy Clustering techniques which, in addition to assigning each observation to one of the identified clusters, also provides a level of probability of belonging to all the available clusters. C-means is a method used to implement these techniques [21].

5. Conclusions

Cluster analysis applied to this CT patient dose database showed that it is possible to effectively synthesize the main currently used exposure conditions and the relative dose indicator distributions, thus highlighting anomalous situations that merit further investigation. With the increasing availability of patient dose data from multicenter sources and wider geographical areas, data mining techniques for the extraction of salient information will be a key element in medical physics in the coming years.

Acknowledgement

This work was supported, in part, by grant from the project “Procedure radiologiche: informazione della dose al paziente” Compagnia San Paolo Torino.

References

[1] Rehani MM. Tracking of examination and dose: overview. Radiat Prot Dosim 2015;165:50–2.
[2] Talati RK, Dunkin J, Parikh S, Moore WH. Current methods of monitoring radiation exposure from CT. J Am Coll Radiol 2013;10:702–7.
[3] Szczykutowicz TP, Bour RK, Rubert N, Wendt G, Pozniak M, Ranallo FN. CT protocol management: simplifying the process by using a master protocol concept. J Appl Clin Med Phys 2015;16:228–43.
[4] Varian HR. Big data: new tricks for econometrics. J Econ Perspect 2014;28:3–28.
[5] McCollough CH. Automated data mining of exposure information for dose management and patient safety initiatives in medical imaging. Radiology 2012;264(2):322–4.
[6] Ikuta I, Sodickson A, Wasser EJ, Warden GI, Gerbaudo VH, Khorasani R. Exposing exposure: enhancing patient safety through automated data mining of nuclear medicine reports for quality assurance and organ dose monitoring. Radiology 2012;264(2):406–13.
[7] Sodickson A, Warden GI, Farkas CE, Ikuta I, Prevedello LM, Andriole KP, et al. Exposing exposure: automated anatomy-specific CT radiation exposure extraction for quality assurance and radiation monitoring. Radiology 2012;264(2):397–405.
[8] Wang S, Pavlicek W, Roberts CC, Langer SG, Zhang M, Hu M, et al. An automated DICOM database capable of arbitrary data mining (including radiation dose indicators) for quality monitoring. J Digit Imaging 2011;24(2):223–33.
[9] Han J, Kamber M, Pei J. Data mining, concepts and techniques. Elsevier; 2012.
[10] Nicol RM, Wayte SC, Bridges AJ, Koller CJ. Experiences of using a commercial dose management system (GE Dose Watch) for CT examinations. Br J Radiol 2016;89:20150617.
[11] Chormunge S, Jena S. Efficiency and effectiveness of clustering algorithms for high dimensional data. Int J Comput Appl 2015;125:35–40.
[12] Odilia Y, Ramdeen KT. The quantitative methods for psychology hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. Quant Methods Psychol 2015;11:8–21.
[13] Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963;58:236–44.
[14] Zambelli AE. A data-driven approach to estimating the number of clusters in hierarchical clustering. F1000Res 2016:5. pii: ISCB Comm J-2809.
[15] Origgi D, Vigorito S, Villa G, Bellomi M, Tosi G. Survey of computed tomography techniques and absorbed dose in Italian hospitals: a comparison between two methods to estimate the dose-length product and the effective dose and to verify fulfilment of the diagnostic reference levels. Eur Radiol 2006;16:227–37.
[16] Kanal KM, Stewart BK, Kolokythas O, Shuman WP. Impact of operator-selected image noise index and reconstruction slice thickness on patient radiation dose in 64-MDCT. Am J Roentgenol 2007;189:219–25.
[17] Tang H, Yu N, Jia Y, Yu Y, Duan H, Han D, et al. Assessment of noise reduction potential and image quality improvement of a new generation adaptive statistical iterative reconstruction (ASIR-V) in chest CT. Br J Radiol 2018;91(1081):20170521.
[18] Matsubara K, Koshida K, Ichikawa K, Suzuki M, Takata T, Yamamoto T, et al. Misoperation of CT automatic tube current modulation systems with inappropriate patient centering: phantom studies. Am J Roentgenol 2009;192(4):862–5.
[19] Serna A, Ramos D, Angosto EG, Garcia-Sanchez AJ, Chans MA, Benedicto-Orovitg JM, et al. Optimization of CT protocols using cause and effect analysis of outliers. Physica Med 2018;55:1–7.
[20] Hummel M, Edelmann D, Kopp-Schneider A. Clustering of samples and variables with mixed-type data. PLoS One 2017;12:e0188274.
[21] Dunn J. A fuzzy relative of the ISODATA process and its use in detecting compact well separated cluster. J Cybern 1974;3(3):32–57.