# Data Visualization in Clinical Practice


**Monique Hendriks, Charalampos Xanthopoulakis, Pieter Vos, Sergio Consoli, and Jacek Kustra**

## 1 Introduction and Related Work

Clinical decision support is an emerging area where the combination of information systems and humans interacts to perform decisions on diagnostics or treatment selection [1, 43]. In this interaction, previously collected data is processed by the system, interfaced to the user, e.g., by means of visualization, and a final decision is made by a human being [26].

In modern information systems, the available information is typically much more than one single individual can interpret within the time constraints that make information—and inferred knowledge—useful for a clinical task [24]. Therefore, trade-offs need to be made on what information is presented and how it is presented to best accomplish the target task.

With the amount of information being overwhelming for a single individual to interpret, we need to limit the amount of information presented to the end-user. Tailoring the presented information to the task at hand, e.g., deciding which treatment is best for a patient, allows for selection of a subset of information useful for that particular task. However, we can never assume that a certain piece of information will not be useful. Hence, there is a trade-off in that potentially useful information may be lost if we limit the amount presented to the end-user too much, while interpretability can be severely compromised if too much information is presented.

M. Hendriks (✉) · C. Xanthopoulakis · P. Vos · S. Consoli · J. Kustra
Philips Research, Eindhoven, The Netherlands
e-mail: monique.hendriks@philips.com; charalampos.xanthopoulakis@philips.com; pieter.vos@philips.com; sergio.consoli@philips.com; jacek.kustra@philips.com



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_11

The manner of presentation also involves a certain trade-off, as there is a wide range of methods for presenting information to an end-user, ranging from tables to risk scores arising from supervised learning methods correlating past data with known outcomes, and visual summaries of data. Different representations may disclose patterns in the data and as such provide the end-user with insights that can influence the final decision. Consider, for example, the case presented in [7]. In the search for a predictive model for death from pneumonia, a neural network and a rule-based model were evaluated. While the neural network was more accurate, the rule-based model was in the end preferred, as it gave more insight into the reasoning of the predictive model. The rule-based model allowed the user of the model to identify possibly useless and even risky relations in the model. In this particular example, a relation was found between presence of asthma as a comorbidity and risk of death, but the relation was not as expected. It was found that having asthma *decreased* the risk of death. This is explained by the fact that patients with asthma presenting with pneumonia were usually admitted directly to the intensive care unit.

Ideally, the information presented to the end-user should be transparent and unbiased. This means that the source of the information should be transparent (how was the raw data manipulated to extract that piece of information) and that any operation that was performed to process the data before displaying it does not introduce bias towards drawing conclusions that may not be valid. Consider, for instance, the case of a mix of continuous and categorical features, such as age and gender; many visual data representation techniques use the distance between feature values. Commonly used distance measures are geometrical, such as the Euclidean distance. Applying a geometrical distance measure to the combination of age and gender with normalized values may lead to a distorted view of the impact of gender compared to age, as the unidimensional distance between "male" and "female" is the extreme value of 1, while the unidimensional distance between two different (normalized) ages is typically much smaller than 1. This inevitably introduces a bias into the data visualization, and it should therefore be made clear to the end-user how the data was processed, so that the user may be aware of this bias.

In this chapter, we organize the sections as follows: In Sect. 2, we explore the added value of flexible visualization methods as compared to validated prediction models, as well as the challenges in data visualization. A data visualization approach that aims at providing ease of interpretability, demonstrating transparency, and reducing inherent bias to a minimum is presented in Sect. 3. We close this chapter with a discussion and conclusion section along with future directions (Sect. 4).

## 2 Motivation

With the widespread adoption of electronic health records (EHRs), patient data storage in clinical practice is becoming digital and standardized. While previously predictive models and guidelines in health care would be developed on data from clinical trials, which are set up to have both strong internal and external validity, now

development of models and guidelines from data from clinical practice becomes possible. This has the advantage that much more data is available and models can be developed more quickly to keep up with the pace of development of better diagnostics and measurements and improvements in treatment. However, the strong requirements on internal and external validation are much more difficult to meet in a clinical practice setting. Therefore, it is important to leverage the expertise of the clinical user to ensure that valid conclusions are drawn, taking into account the uncertainty, while still exploiting the knowledge available from such a large and up-to-date data source.

In the remainder of this section, the practice of modeling from clinical trial data will be evaluated and requirements imposed by the use of clinical practice data will be explored, motivating the choice for investigation of visualization methods for clinical practice data.

From the area of statistics as well as from the area of machine learning, a multitude of methods is available to model data. Given the validity of the design of the trial and the data collection executed in the trial, these methods allow the development, interpretation, and validation of such models. Many of those methods are implemented in modules, packages, or tools readily available on the web (e.g., R [17], SPSS [31], SciPy [28], and Weka [16]). The output generated from these methods typically consists of:

- The model: a structure which may be applied to a new patient, generating a prediction value;
- Training error: a measure of the error of the model in representing the data used to train the model;
- Model performance: a measure of the performance of the model on validation data (not used to train the model).

With some exceptions, these methods typically do not provide any human interpretable description of the model itself. For example, the support vectors provided by the support vector machine (SVM) method can be inspected, but they are not easy to interpret even for a data analytics expert, let alone for a non-expert user of the model. Methods such as decision trees or Bayesian networks do generate visual representations of the model that can be inspected and interpreted by a non-expert user. However, even these simple model representations can quickly become too complex to interpret when the size of the network or decision tree increases or when the number of node relations is high. In health care, data analytics models that outperform treatment guidelines (such as the NCCN guidelines for cancer treatment[1]) often do so because they encompass a larger set of features. For example, in cancer treatment, models outperforming guideline diagnosis and treatment selection often include complex imaging parameters and/or genomic features; see, for example, [38, 45]. Data analytics techniques model the data in a finer granularity than guidelines do. For example, in non-small cell lung cancer

[1] See https://www.nccn.org/, last accessed: 2018-06-14.

staging, the guidelines score tumor size in three categories, smaller than 3 cm, between 3 cm and 7 cm, and larger than 7 cm [13], while a prediction algorithm such as a regression model may take into account the exact tumor size.

The purpose of clinical prediction models usually is to support a doctor in the decision-making process regarding diagnosis or treatment. In the past, such predictive models were typically developed on a large set of patients from clinical trials, ideally from multiple sites, and subsequently validated externally in separate clinical trials, ideally also at multiple sites. Models that are nowadays used in clinical practice, such as the Framingham risk score for coronary heart disease [41], usually have been developed and extensively validated in this manner. They are widely accepted due to this extensive validation.

Data collection in clinical research has always been aimed at data analysis; it is digitized and standardized. As data collection in clinical practice is also becoming digital and standardized, it becomes possible to do additional data analysis on clinical practice data. This allows for types of explorative analysis where it is not necessary to define a hypothesis and the type of data that needs to be collected to test the hypothesis beforehand, as is the case with clinical trials. This in turn allows for earlier insight generation from new data arising, e.g., from new treatments, improvements on devices for imaging, better image analysis techniques, or new diagnostic tests. However, acceptance of such models in practice is more than just a matter of reporting sufficient quality on a validation set. Lack of understanding of a model has been reported as a barrier in adopting a model in clinical practice [20]. Furthermore, less extensively validated models require the doctor to have a better understanding of the limits of the applicability of the model; i.e., the doctor must be able to answer questions such as *What is the level of uncertainty in the predictions?* and *Do the predictions from this model apply in my current context (e.g., using an improved diagnostic imaging device)?*. As medicine is becoming more personalized, the number of features in a model increases, resulting in increasingly complex models. It is therefore important to pay attention to the presentation of a model to the user.

Visualization techniques can help provide more insight into complex models. Visual dominance in humans shows that information processing in the visual domain is much faster and more developed than any other modality [34]. While there is a large variety in data visualization techniques, in general the visual domain allows for more ease of interpretation than, for example, numerical representations of risk scores and confidence intervals. However, even though visual representations may improve ease of interpretation, we should beware that the other requirements are also satisfied. Instilling in the user a sense of awareness of the uncertainty in the data is a challenging task that will trade off against ease of interpretation.

In this chapter, focus will be on visualization techniques that are meant to visualize relations in the data without drawing any inference on, e.g., causality. This should force the user to leverage on his or her own clinical knowledge and to consider the uncertainty in the data. For example, visualization of a dataset may show a strong correlation between tumor size and 2-year survival, but it is still up to the doctor looking at that visualization to conclude whether there is a causal relation

between the two, or whether there may be some other explanation of why they are correlated, such as the difference in treatment between small and large tumors. In that sense, these visualization techniques are related to the philosophy behind unsupervised learning. Unsupervised machine learning is the machine learning task of inferring a function to describe hidden structure from "unlabeled" data (a classification or categorization is not included in the observations). Popular approaches include clustering [11] (e.g., K-means [29], mixture models [3], and hierarchical clustering [37]); anomaly detection [8]; neural networks [35] (e.g., Hebbian learning [21] and generative adversarial networks [39]); approaches for learning latent variable models [12] (e.g., expectation-maximization algorithm [4] and method of moments [18]); blind signal separation techniques [3] (e.g., principal component analysis [19], independent component analysis [9], non-negative matrix factorization [25], and singular value decomposition [2]).

Unsupervised learning techniques exploit correlations in the data, without making any inferences on outcome. As such, unsupervised models provide insights into the data such as which patients are similar or dissimilar to each other, allowing the doctor to make an inference on what is the expected outcome for the patient.

An on-screen display of an unsupervised model is typically done through mapping data points onto a two-dimensional graph, using color and/or shape to indicate which data points are grouped together, e.g., through a dimensionality reduction technique such as principal component analysis (see, e.g., [42]). An example is shown in Fig. 1 [32]. An advantage of such a method is that it exploits methods of processing that humans are very good at. Current research has shown that certain salient features such as color, shape, motion, and spatial position are easily detected and discriminated from each other. In early selection theories of attentional processing, this is termed "preattentive processing." The term refers to a kind of effortless processing for which no attention is needed. Evidence for preattentive processing was found in visual search tasks, where subjects are asked to locate a certain target stimulus among a set of non-target (distracting) stimuli. It

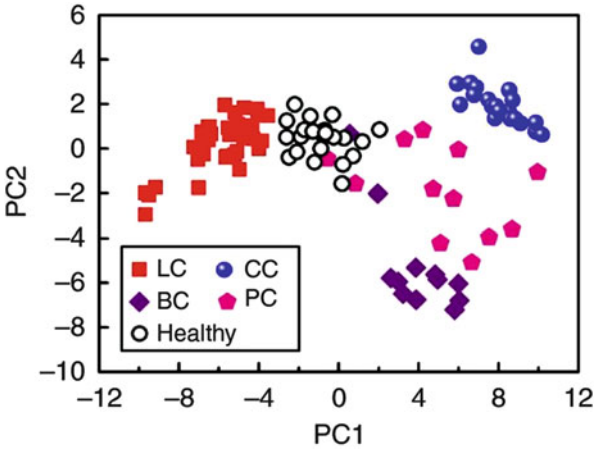


**Fig. 1** An example of a graphical display of clustering [32], demonstrating detection of lung, breast, colorectal, and prostate cancer from exhaled breath using nanosensors

was found that search times for stimuli defined by a single salient feature such as a red shape among green shapes or a circle among squares were much lower than search times for a target stimulus defined by a combination of features such as a red circle among green circles and red and green squares. Search for a single salient feature appears to be effortless; the target subjectively "pops out"[10].

A disadvantage of the type of representation shown in Fig. 1 is that it is difficult to retrace what the feature values of a point are. Knowing the feature values of the groups of patients that belong together is however a strong requirement for helping the user make sense of the clustering. In the next section (Sect. 3), we present data visualization methods accepted for clinical practice that demonstrate correlations and groupings among patients in a dataset while also allowing for inspection of individual feature values.

## 3 Data Visualization Techniques in Clinical Practice

In this section, we provide an example of a visualization technique for decision support accepted for use in clinical practice. It has the aim of selecting the best treatment for a given patient. This is achieved by providing a visual representation comparing patient characteristics to (local) similar patients, who have already been treated.

The parallel coordinates plot is a straightforward and ready to use visualization of multivariate data and has been around for many decades [14]. Figure 2 shows an example of a parallel coordinates plot with patient data. In the parallel coordinates plot, every observation (i.e., a patient) in a dataset is represented with a polyline that crosses a set of parallel vertical axes corresponding to features in the dataset. Parallel coordinates plots readily reveal patients who appear most similar with respect to their characteristics from the "tightness" of their polylines. The competitive advantage of parallel coordinates plots lies in the fact that this tightness can be easily identified in the 2D pattern, while separate multivariate feature values are also still readily recognizable, as opposed to plots derived from dimensionality reduction techniques, such as the one shown in Fig. 1.

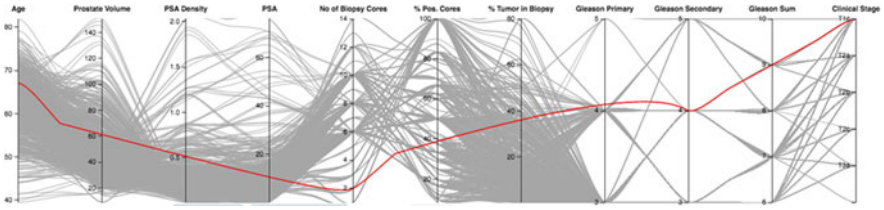


**Fig. 2** Example of a parallel coordinates plot displaying clinical characteristics of prostate cancer patients. The plot displays thousands of patients, represented by polylines. One particular patient currently under observation is represented by a red line

However, the interpretation of parallel plots is dependent on the layout of the parallel coordinates plot. The most important factors are the order of variables and scaling of the axes. The order of variables has an impact on the capability to find relations between the variables; relations between variables that are presented in successive order are more easily seen than relations between variables that are separated from each other by other variable axes that are in between. Furthermore, as the variables are ordered in a linear fashion, a relationship among neighboring variables is implied through the Gestalt principle of proximity (see, e.g., [22]); items that are closer together are perceived as more related than items that are further apart. Proper ordering and selection of the proper subset of variables is therefore essential [44].

Another important factor is the scaling of the variable per axis. Typically, such scaling will be a (linear) normalization such that all axes are of the same length. Consider, for instance, a dataset that contains age and gender. Age typically has a large range of values, while gender only has two unique values. This means that values "male" and "female" will be mapped onto the bottom and the top of an axis that has the same length as the axis which shows age. Furthermore, reversing the values for "male" and "female" results in a different plot. One can also imagine that when a variable has a logarithmic distribution, e.g., many patients have a low blood test value for presence of cancer, mapping to a linear scaled axis will limit the ability to observe patterns.

Parallel coordinates plots become hard to read, when there are many data records included. In the example of Fig. 2, thousands of patients are included, resulting in a vast overlap of lines. This makes it hard to single out sub-populations or to detect patterns in the data. Stratified coloring of the polylines improves the readability and is therefore often applied.

The example of the mapping of age and gender onto an axis in a parallel coordinates plot also makes it clear that parallel coordinates plots display this particular limitation of reduced readability even more so in rendering categorical data. In the example of gender, with just two unique values, all polylines will cross the axis of gender in one of two places.

A data visualization that is better equipped for dealing with categorical data is a parallel sets plot [23]. In the parallel sets technique, the concept of individual lines per patient is substituted for a frequency-based representation. In such a representation, a line represents a subset of patients that have the same categorical feature values. The width of the line is proportional to the size of the subset. See Fig. 3 for an example parallel sets plot based on the Titanic survival data (image generated using R software package Alluvial [5]).

While parallel sets plots are better equipped for dealing with categorical data, they are not suitable for dealing with continuous data. Categorization is therefore often applied as a remedy which may lead to loss of information. Parallel coordinates/sets techniques are therefore limited in use when dealing with heterogeneous data.

Another limitation of parallel sets/parallel coordinates plots is that missing values cause a distortion of the plot. Particularly, in the parallel coordinates plot, a

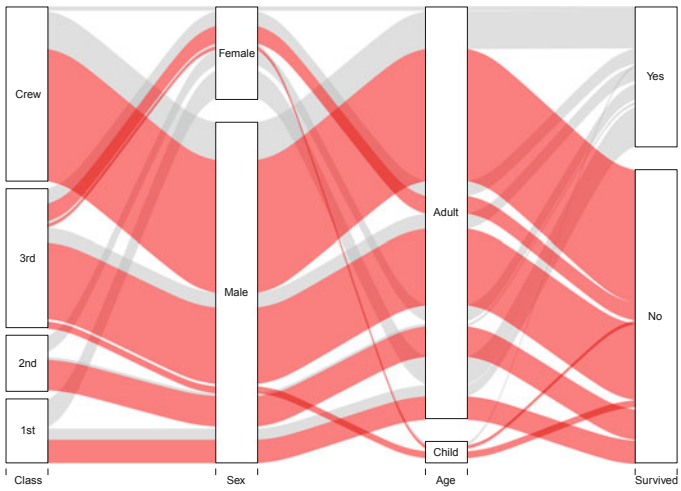


**Fig. 3** Example of a parallel sets plot showing categorical data, where instead of drawing multiple lines, each drawn line represents a different stratification of the Titanic passengers. The width of the line is proportional to the number of passengers

missing value would result in a missing line segment. Research into psychology and attention has shown that humans tend to automatically fill in gaps in a contour [40]. So rendering a line with a missing segment may lead to misleading conclusions regarding the missing values that may not be warranted by the plot itself. The end-user may even be unaware of having made this inference.

Another important consideration from human information processing is that short-term memory generally has a capacity of around 7 (plus or minus 2) items [27]. This means that the number of features that can be included in a parallel coordinates plot such that they can still be reasonably expected to be compared with each other by a user is around 7.

The mentioned limitations are addressed by the circular layout approach described in the next section.

## *3.1 An Extension Towards a Chord Diagram*

Chord diagrams are gaining in popularity for several applications ranging from large software package visualization to visualization of biological data [15]. In the circular layout of a chord diagram, such as provided by Circos,[2] connections between objects or between positions become readily recognizable, while in a linear

[2]Introduction to Circos, Features, and Uses http://www.circos.ca/, last accessed: 2018-01-03.

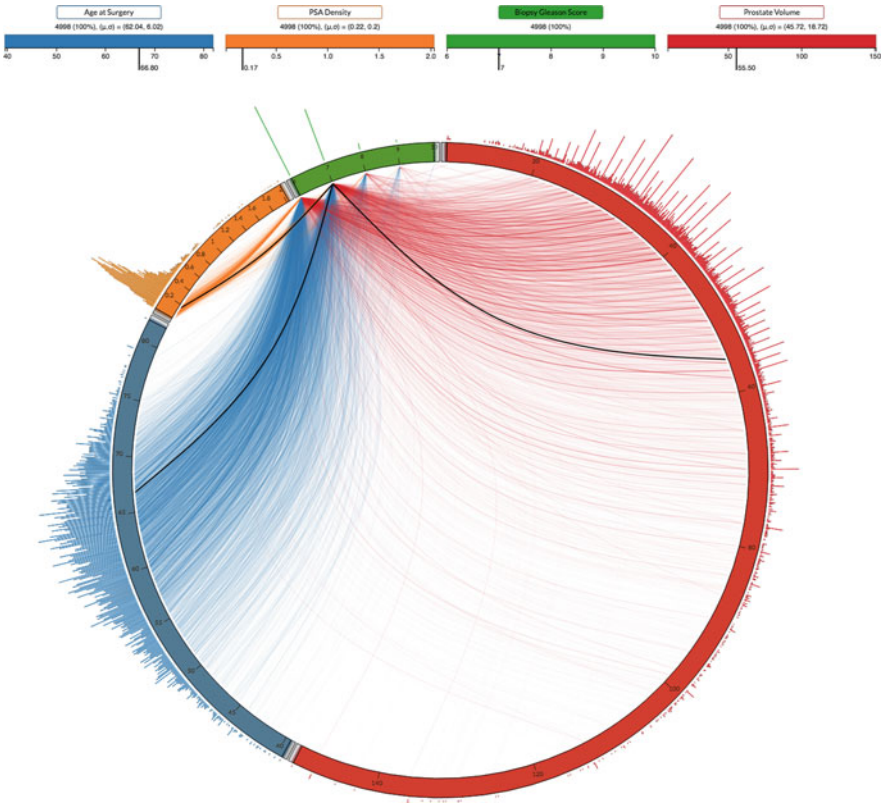


**Fig. 4** A circular layout data visualization of a cohort of prostate cancer patients, showing the clinical parameters age (blue), PSA density (orange), biopsy Gleason score (green), and prostate volume (red) that are commonly used variables to decide which treatment should be provided to the patient

layout, organization of the chart such that multiple connections in a large dataset become easily recognizable is often extremely difficult. It has been shown that pairwise comparisons are efficient in relation-finding [36]. The circular approach exploits this property by connecting pairs of variables. An example of a circular plot with a clinical application is shown in Fig. 4. Here the chord diagram displays prostate cancer patients with the four most prominent variables in the decision-making process of clinicians, i.e., patient age, prostate-specific antigen (PSA) density, biopsy Gleason score, and prostate volume [30].

Note that each colored arc corresponds to a variable. The length of each arc is proportional to the range of values relative to each clinical measure. As such, the extent of each continuous variable domain is mapped to an arc length such that each individual attribute value assumes an equal angle. In this way, outliers are readily recognized.

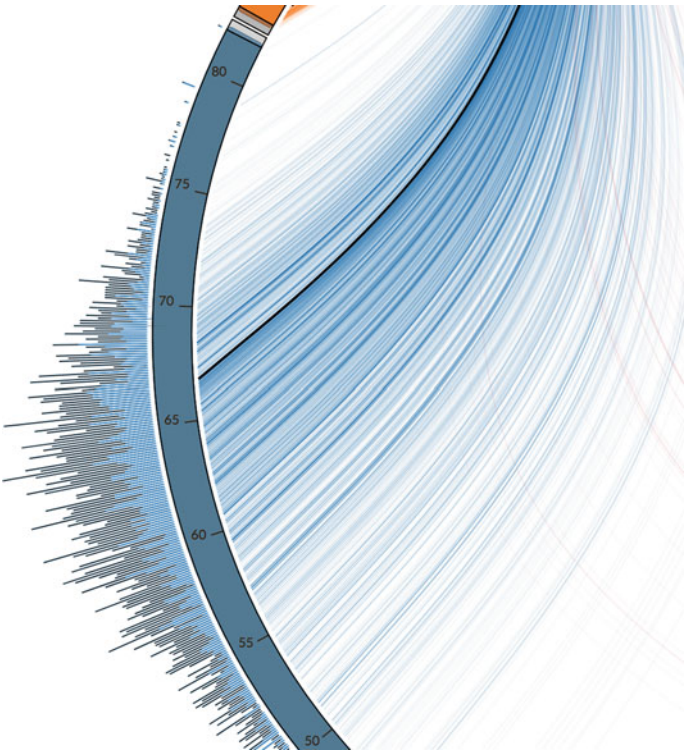


**Fig. 5** Zoom in of the circular plot. Note that the opacity and the thickness of each connecting curve depicts the frequency of occurrence for each tuple

The biopsy Gleason score is an important measure of prostate cancer aggressiveness and is therefore set as the primary measure to which all other features are paired. For each patient, a curve is drawn between the primary measure value and the respective attribute value, i.e., the patient age, the PSA density, and the prostate volume. This promotes the detectability of relations between pairs of clinical measures. Furthermore, opacity and thickness of each connecting curve is used as a means of depicting the frequency of occurrence for each tuple, as shown in Fig. 5. In other words, the more frequently a particular combination of values appears in the dataset, the brighter and wider the curve. Another advantage of the chord diagram is that patients with incomplete data will still be visualized in the figure for pairs of variables that are complete.

The circular layout presented in this chapter also reveals another advantage over parallel coordinates plots: its compact design allows to add several layers of information and detail by adding outer rings. For example, as demonstrated in Fig. 6, a density graph per feature is added to the outside of the ring. This way, clinicians are able to inspect exact feature values of individual patients, as well as the distribution of feature values in one graph, allowing them to draw their own conclusions on the

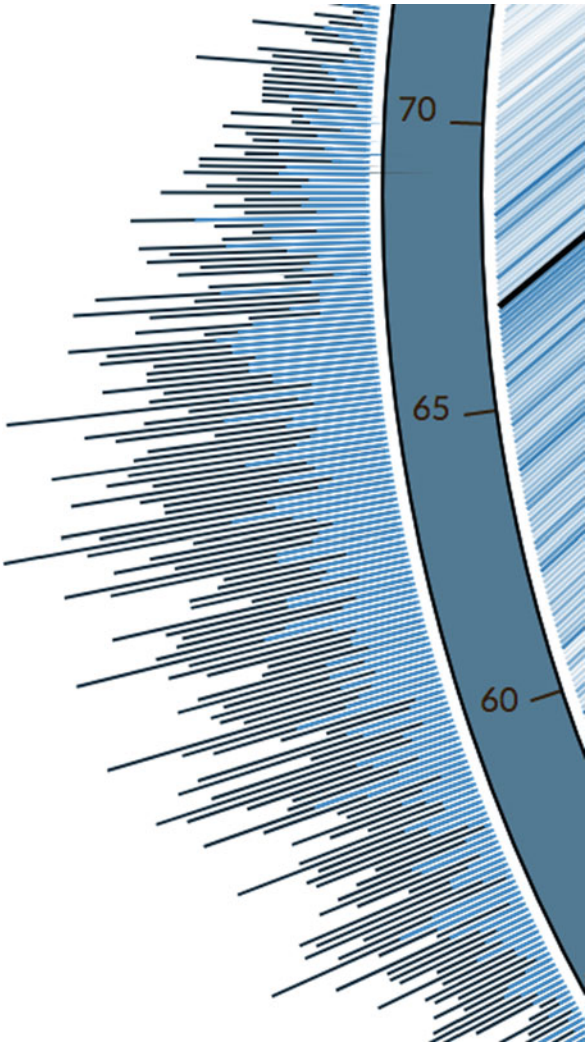


**Fig. 6** Distribution of values along a clinical measure. Note that the count of patients with a certain variable value is displayed as a vertical ray of proportional length perpendicular to the attribute arc. The gray area is the result of filtering of another variable, indicating that this part of the distribution is outside the selected cohort

correlations and variance of respective attributes. Binning of continuous variables is avoided, such that the clinician is in control of evaluating the distribution of variable values to promote unbiased conclusions.

An interactive filtering mechanism is added to the chord diagram by means of brushes alongside each arc. This allows the clinician to select a range of values of interest for a certain variable. The selection results in a subset of patients that match the filtering criteria being highlighted. Such a comparison is also depicted on the distribution of patients alongside each arc, as indicated in Fig. 7. Figures 6 and 5 show the effect of making a selection on a range in one variable on the other variables. In Fig. 7, it can be seen that a range of values for PSA density (orange) is

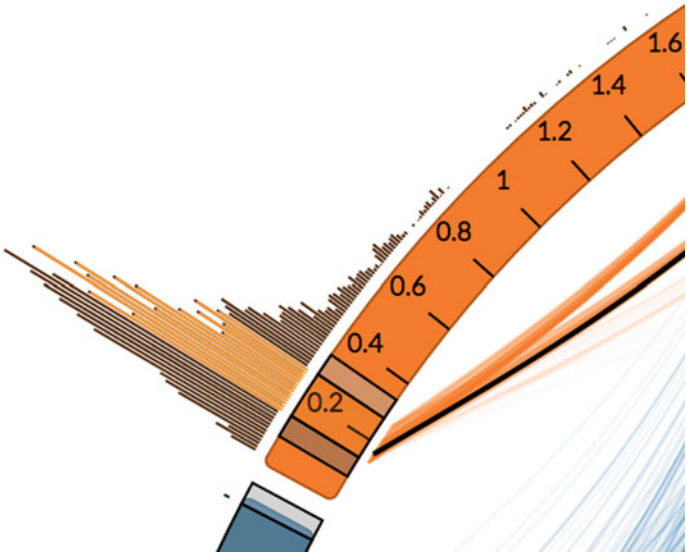


**Fig. 7** Interaction with the circular plot allows for filtering on a specific range of variable values, such that pairs are visualized within the subselection only

selected. In Figs. 6 and 5, the density graph for the variable age (blue) highlights the patients that are within this selected ranges for PSA density, while patients who are outside this range are shown in gray.

This circular approach serves as a means of comparing an individual patient with the population of patients that already have been treated and is well suited for identifying trends and outliers. Figure 8 demonstrates the case of an outlier. The thick black curves refer to a particular patient record with low biopsy Gleason score, low PSA density value, and high prostate volume and a more senior age. Even without the exact numbers depicted on the graph, it is readily recognizable that the patient in question does not fit the general distribution. Upon examining the graph in Fig. 8, clinicians may be prompted to rethink whether these outlier patients should receive the same recommendation for treatment as the general population.

## 4 Discussion and Conclusions

In this chapter, we have discussed the need for more flexible clinical decision support as the fast pace of development of new techniques and treatments causes any extensively validated model to be outdated by the time it is ready for deployment in clinical practice. Data visualization techniques support generation of insight from data without presenting precalculated conclusions to the user. By leaving the

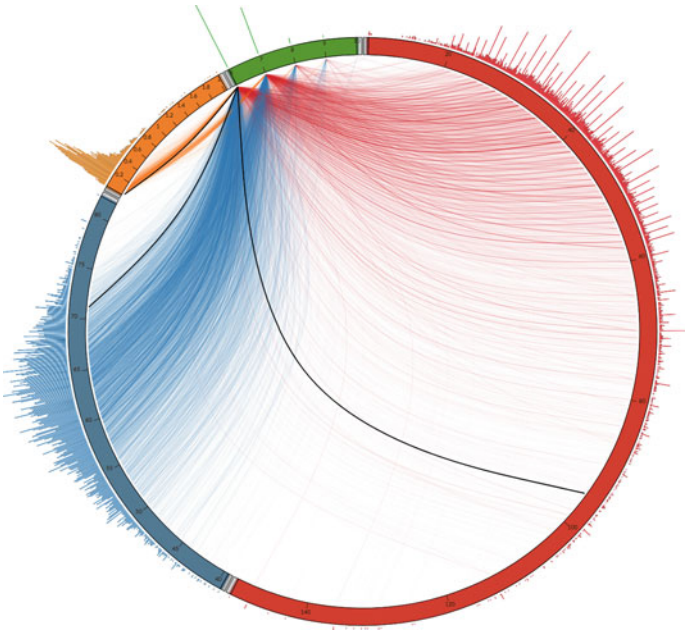

**Fig. 8** Example of a patient under investigation (black line) of which the variable prostate volume (red arc) does not fit the general distribution. This should alert the clinician that this is probably an exceptional patient and care should be taken in the decision-making

decision power in the hands of the human expert, we can provide decision support that is able to keep up with the fast generation of new data.

However, this presents several challenges since, even with the most simple visualization techniques, data is being processed before it is put on the screen and, in that processing, bias may be introduced. Therefore, transparency of which operations were executed on the data to translate it to an on-screen visualization is key. Furthermore, it is important for the user to be aware of the level of uncertainty inherent in the data, as we are sacrificing extensive external validation for flexibility and speed. Finally, leaving the power to draw conclusions in the hands of the doctor also requires ease of interpretation so that the visualization helps the doctor to gain the right insight into the data. Transparency, clarity of the level of uncertainty, and ease of interpretation together should help doctors make informed decisions while staying aware of the risks.

We have discussed that these are not all-or-none end goals in the search for the best possible visualization method; there are trade-offs to be made on the amount of information that is displayed (and the amount that is left out) and the way in which information is presented. We have described how the presented circular approach incorporates these trade-offs. The method offers ease of interpretation through exploitation of the human psychological strength in comparing pairs of features. This may come at a cost of identification of more global patterns among multiple

features, but due to ease of interpretation, it does become possible to incorporate more features compared to any method that focusses more on global patterns. Yet, it is still advised to make a pre-selection of incorporated features through application of clinical domain knowledge, as was done in the example for prostate cancer.

The method is transparent in that it is clear that the range of features corresponds to the length of the arc, the distribution of the data is shown perpendicular to the arc, and the width and brightness of the curves corresponding to the patient data corresponds to the frequency of occurrence. However, it should still be noted that the distances along the arcs can be arbitrarily chosen and particularly the distances between values of categorical features should be carefully interpreted. Integration of the data distribution into the same graph allows for assessment of uncertainty in any conclusions that may be drawn. It can be easily seen how wide the spread is among feature values and whether distributions on a certain feature are skewed to the upper or the lower end.

Future experiments should investigate to what extent the circular approach allows for inclusion of multiple features: how many features can be included without too much loss of ease of interpretation? However, as the amount of data collected is increasing, selective display of information will remain inevitable. This selectivity may be automated, through employing data analytics methods such as clustering or classification to achieve, for example, smart feature selection. However, besides taking away a certain amount of control from the clinician, such automation also comes at the cost of a steeper regulatory path towards incorporation of visualizations in clinical practice.

While selective display will remain an inevitable part of the trade-off between the amount of information displayed and the ease of interpretation, we have shown in this chapter that the trade-off can be softened through choosing the right manner of displaying information. We have shown that a circular approach increases the amount of information we can display without sacrificing ease of interpretation. Additions of solutions such as graph bundling [33] can be explored in the future to allow for even greater increase in the amount of data that can be displayed without sacrificing ease of interpretation.

Finally as the famous quote of George Box explains: “All models are wrong but some are useful” [6]. The more data is collected, the more heterogeneous it will become, thereby inherently requiring a greater amount of simplification and therefore uncertainty in any model we create from that data, be it a machine learning model or a visualization. It therefore becomes important to focus on the second part of the quote and investigate how any model that can still be interpreted by a doctor can be as useful as possible. This requires tuning any model to the correct clinical needs as well as to the strengths and limitations of human information processing.

## References


1. Abernethy, A.P., Etheredge, L.M., Ganz, P.A., Wallace, P., German, R.R., Neti, C., Bach, P.B., Murphy, S.B.: Rapid-learning system for cancer care. J. Clin. Oncol. **28**(27), 4268–4274 (2010)
2. Acharyya, R.: A New Approach for Blind Source Separation of Convolutive Sources. VDM Verlag, Saarbrucken (2008)
3. Alpaydin, E.: Introduction to Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, Cambridge (2009)
4. Assaad, H.E., Same', A., Govaert, G., Aknin, P.: A variational expectation-maximization algorithm for temporal data clustering. Comput. Stat. Data Anal. **103**, 206–228 (2016)
5. Bojanowski, M., Edwards, R.: Alluvial: R Package for Creating Alluvial Diagrams (2016). r package version: 0.1–2 https://github.com/mbojan/alluvial
6. Box, G.E.P.: Science and statistics. J. Am. Stat. Assoc. **71**(356), 791–799 (1976)
7. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721–1730. ACM, New York (2015)
8. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 1–58 (2009)
9. Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. Mach. Learn. **42**, 143–175 (2001)
10. Driver, J.: A selective review of selective attention research from the past century. Br. J. Psychol. **92**(1), 53–78 (2001)
11. Everitt, B.: Cluster Analysis. Wiley, Chichester (2011)
12. Everitt, B.S.: An Introduction to Latent Variables Models. Chapman & Hall/CRC Press, Boca Raton (1984)
13. Goldstraw, P., Crowley, J., Chansky, K., Giroux, D.J., Groome, P.A., Rami-Porta, R., Postmus, P.E., Rusch, V., Sobin, L., for the Study of Lung Cancer International Staging Committee IA, et al.: The IASLC lung cancer staging project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the tnm classification of malignant tumours. J. Thoracic Oncol. **2**(8), 706–714 (2007)
14. Heinrich, J., Weiskopf, D.: State of the art of parallel coordinates. In: Eurographics (STARs), pp. 95–116 (2013)
15. Hinich, V., Vaintrob, A.: Cyclic operads and algebra of chord diagrams. Sel. Math. **8**(2), 237–282 (2002)
16. Holmes, G., Donkin, A., Witten, I.H.: Weka: a machine learning workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, pp. 357–361. IEEE, Piscataway (1994)
17. Ihaka, R., Gentleman, R.: R: a language for data analysis and graphics. J. Comput. Graph. Stat. **5**(3), 299–314 (1996)
18. Jesus, J., Chandler, R.E.: Estimating functions and the generalized method of moments. Interface Focus **1**(6), 871–885 (2011)
19. Jolliffe, I.: Principal Component Analysis, 2nd edn. Springer, New York (2002)
20. Kang, J., Schwartz, R., Flickinger, J., Beriwal, S.: Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. Int. J. Radiat. Oncol. Biol. Phys. **93**(5), 1127–1135 (2015)
21. Keysers, C., Perrett, D.I.: Demystifying social cognition: a Hebbian perspective. Trends Cogn. Sci. **8**(11), 501–507 (2004)
22. Koffka, K.: Principles of Gestalt Psychology, vol. 44. Routledge, Abingdon (2013)
23. Kosara, R., Bendix, F., Hauser, H.: Parallel sets: interactive exploration and visual analysis of categorical data. IEEE Trans. Vis. Comput. Graph. **12**(4), 558–568 (2006)
24. Krumholz, H.M.: Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff. **33**(7), 1163–1170 (2014)

25. Li, L., Zhang, Y.J.: Survey on algorithms of non-negative matrix factorization. Tien Tzu Hsueh Pao/Acta Electron. Sin. **36**(4), 737–743 (2008)
26. Liang, W., Zhang, L., Jiang, G., Wang, Q., Liu, L., Liu, D., Wang, Z., Zhu, Z., Deng, Q., Xiong, X., Shao, W., Shi, X., He, J.: Development and validation of a nomogram for predicting survival in patients with resected non-small-cell lung cancer. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. **33**(8), 861–869 (2015)
27. Miller, G.A.: The magic number seven plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**, 91–97 (1956)
28. Millman, K.J., Aivazis, M.: Python for scientists and engineers. Comput. Sci. Eng. **13**(2), 9–12 (2011)
29. Modha, D.S., Spangler, W.S.: Feature weighting in k-means clustering. J. Mach. Learn. **52**, 217–237 (2001)
30. Mottet, N., Bellmunt, J., Bolla, M., Briers, E., Cumberbatch, M.G., De Santis, M., Fossati, N., Gross, T., Henry, A.M., Joniau, S., et al.: Eau-estro-siog guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent. Eur. Urol. **71**(4), 618–629 (2017)
31. Nie, N.H., Bent, D.H., Hull, C.H.: SPSS: statistical package for the social sciences. Tech. rep., McGraw-Hill, New York (1970)
32. Peng, G., Hakim, M., Broza, Y.Y., Billan, S., Abdah-Bortnyak, R., Kuten, A., Tisch, U., Haick, H.: Detection of lung, breast, colorectal, and prostate cancers from exhaled breath using a single array of nanosensors. Br. J. Cancer **103**(4), 542 (2010)
33. Peysakhovich, V., Hurter, C., Telea, A.: Attribute-driven edge bundling for general graphs with applications in trail analysis. In: 2015 IEEE Pacific Visualization Symposium, PacificVis 2015, Hangzhou, 14–17 April, pp. 39–46 (2015)
34. Posner, M.I., Nissen, M.J., Klein, R.M.: Visual dominance: an information-processing account of its origins and significance. Psychol. Rev. **83**(2), 157 (1976)
35. Ripley, B.: Pattern Recognition and Neural Networks. Cambridge University Press, New York (2007)
36. Saaty, T.L.: Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process. Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales Serie A Matematicas **102**(2), 251–318 (2008)
37. Trevor, R.T., Friedman, J.: The Elements of Statistical Learning: Data mining, Inference, and Prediction. Springer, New York (2009)
38. Wang, L., Mullerad, M., Chen, H.N., Eberhardt, S.C., Kattan, M.W., Scardino, P.T., Hricak, H.: Prostate cancer: incremental value of endorectal MR imaging findings for prediction of extracapsular extension. Radiology **232**(1), 133–139 (2004)
39. Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., Wang, F.Y.: Generative adversarial networks: introduction and outlook. IEEE/CAA J. Autom. Sin. **4**(4), 588–598 (2017)
40. Wickens, C., Lee, J., Liu, Y., Gordon-Becker, S.E.: Designing for People: An Introduction to Human Factors Engineering, 3rd edn. CreateSpace, Charleston (2018)
41. Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., Kannel, W.B.: Prediction of coronary heart disease using risk factor categories. Circulation **97**(18), 1837–1847 (1998)
42. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemom. Intell. Lab. Syst. **2**(1–3), 37–52 (1987)
43. Wu, B., Ricchetti, F., Sanguineti, G., Kazhdan, M., Simari, P., Jacques, R., Taylor, R., McNutt, T.: Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning. Int. J. Radiat. Oncol. Biol. Phys. **79**(4), 1241–1247 (2011)
44. Yang, J., Peng, W., Ward, M.O., Rundensteiner, E.A.: Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In: IEEE Symposium on Information Visualization, INFOVIS 2003, pp. 105–112. IEEE, Piscataway (2003)
45. Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J.E., Song, C., Gutman, D.A., Halani, S.H., Vega, J.E.V., Brat, D.J., et al.: Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci. Rep. **7**(1), 11707 (2017)