How does Radiomics actually work? – Review

Wie geht Radiomics eigentlich? – Review

Authors

Ulrike Irmgard Attenberger[1], Georg Langs[2]

Affiliations

1 Department of Diagnostic and Interventional Radiology, Medical Faculty, University Hospital Bonn, Germany
2 Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab, Medical University of Vienna, Wien, Austria

Key words

machine learning, radiomics, artificial intelligence, features

received 02.03.2020
accepted 05.10.2020
published online 02.12.2020

Bibliography

Fortschr Röntgenstr 2021; 193: 652–657
DOI 10.1055/a-1293-8953
ISSN 1438-9029

Georg Thieme Verlag KG, Rüdigerstraße 14,
70469 Stuttgart, Germany

Correspondence

Prof. Ulrike Irmgard Attenberger
Department of Diagnostic and Interventional Radiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany
Tel.: +49/2 28/28 71 58 71
ulrike.attenberger@ukbonn.de

ABSTRACT

Personalized precision medicine requires highly accurate diagnostics. While radiological research has focused on scanner and sequence technologies in recent decades, applications of artificial intelligence are increasingly attracting scientific interest as they could substantially expand the possibility of objective quantification and diagnostic or prognostic use of image information.

In this context, the term “radiomics” describes the extraction of quantitative features from imaging data such as those obtained from computed tomography or magnetic resonance imaging examinations. These features are associated with predictive goals such as diagnosis or prognosis using machine learning models. It is believed that the integrative assessment of the feature patterns thus obtained, in combination with clinical, molecular and genetic data, can enable a more accurate characterization of the pathophysiology of diseases and more precise prediction of therapy response and outcome.

This review describes the classical radiomics approach and discusses the existing very large variability of approaches. Finally, it outlines the research directions in which the interdisciplinary field of radiology and computer science is moving, characterized by increasingly close collaborations and the need for new educational concepts. The aim is to provide a basis for responsible and comprehensible handling of the data and analytical methods used.

Key points:

- Radiomics is playing an increasingly important role in imaging research.
- Radiomics has great potential to meet the requirements of precision medicine.
- Radiomics analysis is still subject to great variability.
- There is a need for quality-assured application of radiomics in medicine.

Citation Format

- Attenberger UI, Langs G, . How does Radiomics actually work? – Review. Fortschr Röntgenstr 2021; 193: 652–657

ZUSAMMENFASSUNG

Personalisierte Präzisionsmedizin setzt eine hochakkurate Diagnostik voraus. Während die radiologische Forschung sich in den letzten Jahrzehnten mit Scanner- und Sequenztechnologien beschäftigt hat, rücken zunehmend Anwendungen der künstlichen Intelligenz in das wissenschaftliche Interesse, da sie die Möglichkeit der objektiven Quantifizierung und diagnostischen bzw. prognostischen Nutzung von Bildinformationen substanziell erweitern könnten.

In diesem Zusammenhang beschreibt der Begriff „Radiomics“ die Extraktion quantitativer Merkmale aus Bilddaten wie zum Beispiel von Computertomografie- oder Magnetresonanztomografie-Untersuchungen. Diese Merkmale werden mithilfe von Modellen des maschinellen Lernens mit Vorhersagezielen wie Diagnose oder Prognose in Zusammenhang gebracht. Man geht davon aus, dass die integrative Beurteilung der so erhobenen Merkmalsmuster in Verbindung mit klinischen, molekularen und genetischen Daten eine genauere Charakterisierung der Pathophysiologie von Erkrankungen sowie eine präzisere Vorhersage von Therapieansprechen und Outcome ermöglichen kann.

In dieser Übersichtsarbeit werden der klassische Radiomics-Ansatz beschrieben und die bestehende sehr große Variabilität an Zugängen diskutiert. Abschließend werden Forschungsrichtungen skizziert, in die sich das von zunehmend enger Kollaboration zwischen Radiologie und Computerwissenschaften und der Notwendigkeit neuer Ausbildungskonzepte gekennzeichnete interdisziplinäre Feld bewegt. Ziel ist es, eine Grundlage für verantwortungsvollen, nachvollziehbaren Umgang mit eingebrachten Daten und angewandten Analysemethoden zu ermöglichen.

Introduction

The demands of personalized precision medicine require highly accurate diagnostics. Although in recent decades radiological research has focused on the evaluation of scanner and sequence technologies for more accurate disease diagnosis, scientific interest is now focused on current implementations of artificial intelligence (AI) for optimized diagnostics. The implementation possibilities for AI techniques in radiology are manifold: automated lesion detection and characterization, creation of biobanks, dose optimization, structured reporting and radiomics [2, 3]. For the sake of completeness, it should not be forgotten that AI techniques are also used in the latest generation of scanners to optimize data acquisition itself.

The term "radiomics" describes the extraction of quantitative features from image data such as examinations using computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET) and correlation with clinical, genetic or molecular data using AI methods such as machine learning or deep learning. The concept appears promising: using AI methods, information can be extracted from image data that goes far beyond what the human eye can detect. It is assumed that the assessment of these characteristics and feature patterns obtained from the image data when combined with clinical, molecular and genetic data can enable a more precise characterization of the pathophysiology of diseases as well as a statement on therapy response and probable outcome. Some of the applied techniques have been known for decades, but have been developed substantially in recent years, opening up new approaches to the automated exploitation of image information. Publications on this topic go back to the end of the 1940 s, and models such as neural networks were also intensively researched in the 1980 s [4]. Optimized computing power together with methodological advances and increasing availability of large amounts of data to facilitate the training of models have led to a resumption of this work with impressive results [5], resulting in a more timely and efficient utilization of these techniques – a basis for subsequent potential clinical implementation. The scope of application in imaging diagnostics is diverse and ranges from oncological to cardiac and musculo-skeletal diagnostics.

Radiomics is playing an increasingly important role in imaging research due to its great potential to meet the requirements of precision medicine. Numerous studies provide an overview of the underlying concepts [6, 7]. However, it should be noted that every single step of radiomics analysis is subject to great variability. A responsible, comprehensible handling of the submitted data and applied analysis methods is therefore an indispensable basic requirement. Due to the novel way of dealing with image data, an even closer collaboration with medical imaging computing data scientists will be required in the future, as well a restructuring of radiological training.

Radiomics, which describes a subset of AI implementation possibilities in radiology, follows an explicit scheme according to which image data is processed, segmented and analyzed. This overview article will present and explain this analysis.

Radiomics Hands-on

The 6 Phases of Radiomics Analysis

A radiomics analysis can essentially be divided into 6 steps: (i) data acquisition, (ii) definition of a region of interest (ROI), (iii) data (pre) processing, (iv) feature extraction, (v) selection of the features relevant to the problem and (vi) classification (▶ **Fig. 1**) [8].

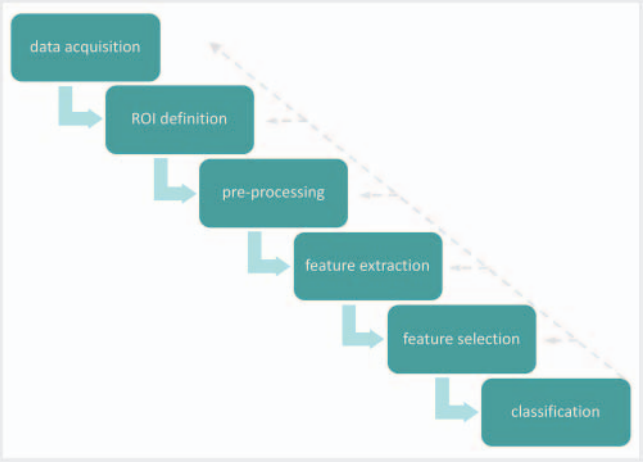
Data Acquisition

The way in which the data are acquired has a significant influence on the result of the radiomics analysis; therefore it is desirable to use imaging protocols that are standardized, reproducible and comparable [9]. For example, a study by Waugh et al. showed that a higher time-to-repetition (TR) enables better discrimination of texture features in breast MRI [10]. In their publication, Baessler et al. systematically tested the factors influencing the choice of sequence in MRI on feature robustness [11]. A high-resolution FLAIR sequence provided the highest feature robustness. On the other hand, the T2-weighted sequence with lower resolution acquired in comparison achieved the poorest feature stability. There were also differences in robustness among the various feature groups (matrices). The shape and GLZLM (GrayLevel Zone Length Matrix) groups achieved the highest robustness, while the histogram-based features were the least robust [11]. For this reason, Lambin et al. call for a stratified approach to data selection: detailed disclosure of imaging sequences, robust segmentation e. g. by multiple evaluators, phantom studies and imaging at different time points [9].

ROI Definition

Optimal ROI Size and Feature Maps

After data acquisition, the region of interest (ROI) is defined, which describes the area over which further analysis will occur. Most of the work related to radiomics deals with issues in oncology, and the ROI is typically set to identify the location of a lesion and apply the subsequent analysis accordingly. Here, too, there is great variability in the methodology of the ROI definition, which in turn has significant influence on the result. Three different ROI



▶ **Fig. 1** The 6 phases of a radiomics analysis Depending on the intermediate or final results, some or all of the analytical steps may have to be repeated.

variants can be selected: an ROI that follows the contour of the lesion, one that surrounds the entire lesion at right angles (bounding box), and a partial ROI drawn in the center of a section of the lesion [8]. Although bounding boxes are easier to create, and are often sufficient, precise segmentation of lesions, evaluation of shape, and more accurate analysis of contrast at the lesion edges support a better understanding of the lesion. In addition to ROI shape and placement, the ROI size also has a significant impact on the result. Sikiö et al. demonstrated a correlation between ROI size and feature stability [12]. Using a spatial resolution of $0.5 \times 0.7\ mm^2$ and a slice thickness of 4 mm, feature stability was lowest with an 80 × 80 pixel ROI; the most stable features were achieved with an ROI of 180 × 180 pixels [12].

Segmentation Methods

Segmentation has two tasks: 1) it can make the analysis more specific by allowing explicit access within or outside a lesion; 2) the shape of the segmented lesion itself is a relevant source of features revealed by segmentation. The segmentation of structures in medical image data is an intensively researched field, and accordingly yields different possibilities. Manual segmentation is complemented by automated segmentation methods using special algorithms such as region-growing, level sets for even structures, or, most recently, successful deep learning methods such as so-called U-nets [13, 14]. To date, manual segmentation using an expert reader has been considered the gold standard [15]. However, inter-reader reliability, the reproducibility of the segmentation performed and time required to manually segment large amounts of data are problematic [16, 17]. To reduce this bias, Lambin et al. recommend multiple segmentation, multi-reader analysis, exclusion of high noise segmentation and the use of data from different breathing cycles [9]. In principle, depending on the available data, segmentation can be performed in both 2D and 3D image data. While 2D analysis allows less differentiation in shape and lesions, it is more independent of often highly variable imaging parameters such as slice thickness.

Image Processing and Preprocessing

Image preprocessing precedes the actual feature extraction. Depending on the data set, this includes interpolation, signal normalization and gray value reduction.

Interpolation of the image data allows standardization of the spatial resolution in data. Studies have shown that higher resolution allows optimized feature extraction. In a study by Mayerhoefer et al., the best results were obtained with interpolation factors of 2–4 [1]. Comparability of the features obtained in the analysis is relevant for signal normalization. Three main approaches are described in the literature: min/max, the Z score and mean ± 3σ [18]. The "mean ± 3σ" method means that the intensities are normalized within μ± 3σ, where μ describes the mean value of the gray values within the ROI, and σ the standard deviation. Consequently, gray values that are outside the range [μ – 3σ, μ + 3σ] are not considered for the analysis.

The reduction of gray values in the form of so-called "binning" during feature extraction results in an improvement in the signal-to-noise (SNR) ratio. It maps the gray value range occurring in the image as frequency distributions. Gray values used in the literature are 16, 32, 64, 128 and 256. In their study, Chen et al. recommend using 32 gray values [19], whereas Mahmoud-Ghonheim et al. use 128 [20].

Image preprocessing has a significant influence on feature robustness. Using a phantom, Wichtmann et al. systematically investigated the influence of spatial resolution, gray value reduction and signal normalization on feature robustness [21]. They demonstrated that only 4 features, skewness (histogram), volume [ml] (shape), volume [vox] (shape) and run length non-uniformity [RLNU] (Gray Level Run Length Matrix, GLRLM), RLNU (GLRLM), remained robust over the variation of all parameters.

This clearly shows that specific recommendations for image processing are necessary.

Feature Extraction

Features that are typically used for radiomics analyses can be divided into 4 primary groups: First Order Statistics, Shape, and Texture Features, as well as Features obtained by wavelet transformation of relevant image sections [16]. The group of Texture Feature matrices include the matrices Gray Level Co-occurence Matrix (GLCM), GLRLM, Gray Level Size Zone Matrix (GLSZM), Gray Level Dependence Matrix (GLDM) and Neighboring Grey Tone Difference Matrix (NGTDM). Multiple features are subsumed under each of these matrices. It should be noted that there is great variation in nomenclature, methodology and software implementation [22]. ▶ **Table 1** provides a typical overview of the features of the individual matrices [23]. At the same time, efforts are being made towards the invariance of features with respect to protocols and corresponding standardization efforts [24]. The selection of feature extractors is based on the expectation of which characteristics are relevant for the analysis, and accordingly, extractors are often chosen or constructed that are invariant to, for example, global rotation or very low frequency gray value changes.

Baessler et al. impressively demonstrated the diagnostic value of texture features for the diagnosis of myocarditis using MRI.

▶ **Table 1** Overview of the features of the individual matrices [21].

first order statistics features	shape and size based features	textural features	wavelet features
		grey-level co-occurence matrix based	
energy	compactness	autocorrelation	
entropy	maximum 3D Diameter	cluster prominence	
kurtosis	spherical disproportion	gray-level run-length matrix based	
maximum	sphericity	gray level non uniformity	
mean	surface area	run length non uniformity	

Their study showed that texture features were able to differentiate patients with biopsy-proven myocarditis from a healthy control group in the same way as conventional MRI parameters. However, unlike the texture features, the conventional MRI parameters did not allow differentiation between a healthy control group and patients with negative biopsy but clinical suspicion of myocarditis. There was only a statically significant difference for the texture features, especially RLNU and Gray Level Non-Uniformity [25]. Radiomics allowed a more precise diagnostic differentiation between patients with myocarditis and healthy controls compared to the current standard.

Feature Selection

A major problem in radiomics analysis is the risk of overfitting the data, which occurs especially when the number of features exceeds the number of records, thus severely limiting the meaningfulness of the analysis. Overfitting can be avoided by reducing dimensionality, i. e. by selecting features to be used for analysis and prediction. This can be based on two foundations: features that are reproducible, robust, and non-redundant can be selected without knowledge of the target issue and allow feature reduction without bias [8, 16]. Feature selection based on how "informative" i. e. relevant, a feature is in the sense of the issue is an effective strategy, but also carries the risk of overfitting. Methods developed from machine learning such as random forests allow an effective selection of informative features while providing robustness against large amounts of non-informative features [26]. In this case, however, as described below, an evaluation of the ultimately resulting predictive accuracy on an independent validation data set that was neither used to train the model nor to select the features is essential [27].

Test-re-test data sets can be used to assess the stability of features, and only those stable features are then used for further analysis. The concordance correlation coefficient (CCC), the dynamic range (DR) and the correlation coefficient across all samples are suitable for testing robustness and reproducibility. Studies have shown that the number of features can thus be reduced considerably, e. g. from 397 to 39 [16]. Furthermore, intra- and inter-observer variability can be tested using the intraclass correlation coefficient (ICC) and Bland-Altmann plots. In addition to the statistical approaches listed here, machine learning methods such as random forests can also be used to identify relevant features for resolving the issue, e. g. the differentiation of benign/malignant.

Classification/Modeling

In addition to the statistical approaches listed here, supervised learning approaches are currently most widespread, i. e. a machine is instructed using training data sets with knowledge of the input vector (features) and the output value (target). After this training the thus developed algorithm is applied to a test data set. At this point the extracted characteristics are used for prediction, whereby a key property of relevant methods such as support vector machines or random forests is that they not only evaluate the relationship between isolated features and the prediction target, but can exploit feature groups as *multivariate patterns*. At this juncture, very rapid progress is also underway, which, enabled by deep learning techniques, increasingly combines the construction of features, their selection and prediction into common models.

Validation

The final step is corroboration using a validation data set. The predictive performance of the algorithm is tested using ROC/AUC (receiver operating characteristic/area under the curve) analysis [28]. The separation between data used for the training or development of the prediction models and selection of features and those used as validation data is essential. This is necessary to ensure an overly optimistic assessment of the forecast accuracy. As a middle course, cross-validation can be used, in which the training and test data set are iteratively separated. It must be taken into account that the respective test data could be the basis for modeling decisions and therefore do not allow a completely independent assessment – a separate validation data set is required for this.

Parmar et al. have tested the stability and predictive performance of different feature selection and classifier methods [28]. Their results showed that among the different feature selection methods, the Wilcoxon test-based method (WLCX) and mutual information maximization (MIM) achieved the highest stability. Among the classifiers, Bayesian achieved the best performance with an AUC value of 0.64 (SD± 0.05).

Due to the great variability of radiomics analysis, standardization of data collection, evaluation criteria and reporting is necessary. To this end Lambin et al. have defined a "Radiomics Quality Score" (RQS) [9], which describes a standardized analytical process starting with data selection, through imaging, feature extraction, analysis and modeling, as well as report generation. Each of these steps is divided into several sub-steps for which there are scoring points. The maximum achievable score (total RQS) is 36. The definition and introduction of an RQS is an essential step to-

wards a quality-assured application of radiomics in medicine, which aims to counter the variability problem of analysis – which already begins with the primary image data acquisition – by a dedicated reporting of the individual steps. The introduction of an RQS score seems particularly relevant in view of the expected future connection of clinical decision support systems with radiomic data [9].

Where does this lead?

In addition to *standard* radiomics approaches that use predefined features, recent development in the field of deep learning, the possibility to combine feature design and predictive model training and to implement them with effective model architectures, plays an increasingly important role in the use of complex image data [7, 29]. On the one hand, this enables the use of image information that is not covered by traditional features. On the other hand, there is the problem of interpreting deep learning models, the solution of which is increasingly the focus of research [30].

Summary

Radiomics is playing an increasingly important role in medical imaging due to its great potential to meet the requirements of precision medicine. However, it should be noted that every single step of radiomics analysis is subject to great variability. A responsible, comprehensible handling of the submitted data is therefore an indispensable basic requirement. In the future, radiomics will require an even closer collaboration with medical imaging computing data scientists, as well a restructuring of radiological training.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Mayerhoefer ME, Szomolanyi P, Jirak D et al. Effects of magnetic resonance image interpolation on the results of texture-based pattern classification: a phantom study. Invest Radiol 2009; 44: 405–411. doi:10.1097/RLI.0b013e3181a50a66

[2] European Society of R. What the radiologist should know about artificial intelligence – an ESR white paper. Insights Imaging 2019; 10: 44 doi:10.1186/s13244-019-0738-2

[3] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016; 278: 563–577. doi:10.1148/radiol.2015151169

[4] LeCun Y, Boser B, Denker JS et al. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation 1989; 1: 541–551. doi:10.1162/neco.1989.1.4.541

[5] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Lake Tahoe, Nevada: Curran Associates Inc. 2012

[6] Mayerhoefer ME, Materka A, Langs G et al. Introduction to Radiomics. J Nucl Med 2020; 61: 488–495. doi:10.2967/jnumed.118.222893

[7] Langs G, Rohrich S, Hofmanninger J et al. Machine learning: from radiomics to discovery and routine. Radiologe 2018; 58: 1–6. doi:10.1007/s00117-018-0407-3

[8] Larroza A, Bodí V, Moratal D. Texture analysis in magnetic resonance imaging: review and considerations for future applications. Assessment of cellular and organ function and dysfunction using direct and derived MRI methodologies. 2016: 75–106

[9] Lambin P, Leijenaar RTH, Deist TM et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology 2017; 14: 749–762. doi:10.1038/nrclinonc.2017.141

[10] Waugh SA, Lerski RA, Bidaut L et al. The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms. Med Phys 2011; 38: 5058–5066. doi:10.1118/1.3622605

[11] Baessler B, Weiss K, Pinto Dos Santos D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. Invest Radiol 2019; 54: 221–228. doi:10.1097/RLI.0000000000000530

[12] Sikiö M, Holli-Helenius KK, Ryymin P et al. The effect of region of interest size on textural parameters. In, 2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA): IEEE 2015: 149–153

[13] Brox T, Bruhn A, Weickert J. Variational motion segmentation with level sets. In, European Conference on Computer Vision: Springer 2006: 471–483

[14] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In, International Conference on Medical image computing and computer-assisted intervention: Springer 2015: 234–241

[15] Raykar VC, Yu S, Zhao LH et al. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In, Proceedings of the 26th Annual international conference on machine learning 2009: 889–896

[16] Kumar V, Gu Y, Basu S et al. Radiomics: the process and the challenges. Magn Reson Imaging 2012; 30: 1234–1248. doi:10.1016/j.mri.2012.06.010

[17] Watadani T, Sakai F, Johkoh T et al. Interobserver variability in the CT assessment of honeycombing in the lungs. Radiology 2013; 266: 936–944. doi:10.1148/radiol.12112516

[18] Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging 2004; 22: 81–91. doi:10.1016/j.mri.2003.09.001

[19] Chen W, Giger ML, Li H et al. Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. Magn Reson Med 2007; 58: 562–571. doi:10.1002/mrm.21347

[20] Mahmoud-Ghoneim D, Alkaabi MK, de Certaines JD et al. The impact of image dynamic range on texture classification of brain white matter. BMC Med Imaging 2008; 8: 18 doi:10.1186/1471-2342-8-18

[21] Wichtmann BD, Attenberger UI, Harder FN et al. Influence of image processing on the robustness of radiomic features derived from magnetic resonance imaging – a phantom study. 2019 ps://submissions.mirasmart.com/ISMRM2019/ViewSubmission.aspx?sbmID=7334

[22] Aerts HJ, Velazquez ER, Leijenaar RT et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications 2014; 5: 1–9

[23] Aerts HJ. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. JAMA Oncol 2016; 2: 1636–1642. doi:10.1001/jamaoncol.2016.2631

[24] Andrearczyk V, Depeursinge A, Müller H. Learning cross-protocol radiomics and deep feature standardization from CT images of texture phantoms. In, Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications: International Society for Optics and Photonics 2019: 109540I

[25] Baessler B, Luecke C, Lurz J et al. Cardiac MRI Texture Analysis of T1 and T2 Maps in Patients with Infarctlike Acute Myocarditis. Radiology 2018; 289: 357–365. doi:10.1148/radiol.2018180411

[26] Breiman L. Random forests. Machine Learning 2001; 45: 5–32. doi:10.1023/A:1010933404324

[27] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research 2003; 3: 1157–1182

[28] Parmar C, Grossmann P, Bussink J et al. Machine Learning methods for Quantitative Radiomic Biomarkers. Sci Rep 2015; 5: 13087 doi:10.1038/srep13087

[29] Litjens G, Kooi T, Bejnordi BE et al. A survey on deep learning in medical image analysis. Med Image Anal 2017; 42: 60–88. doi:10.1016/j.media.2017.07.005

[30] Holzinger A, Langs G, Denk H et al. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery 2019; 9: e1312 doi:10.1002/widm.1312