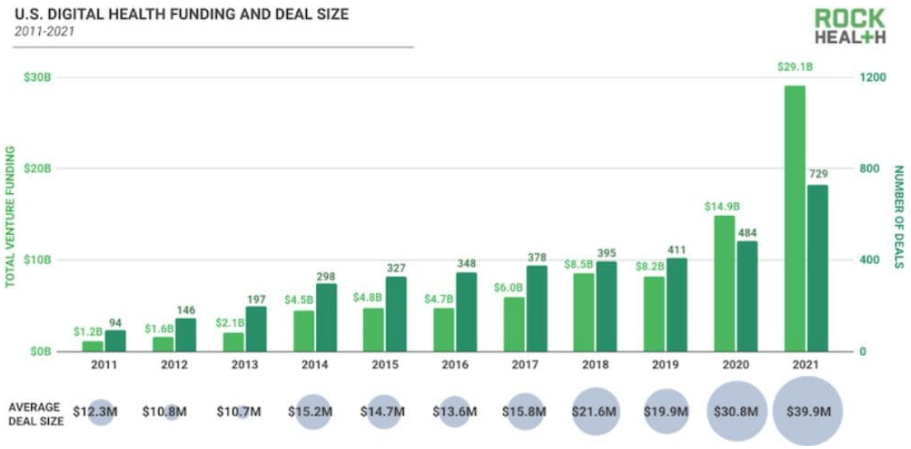
Cancerworld

AI application in diagnosis

Adriana Albini / 11 January 2023

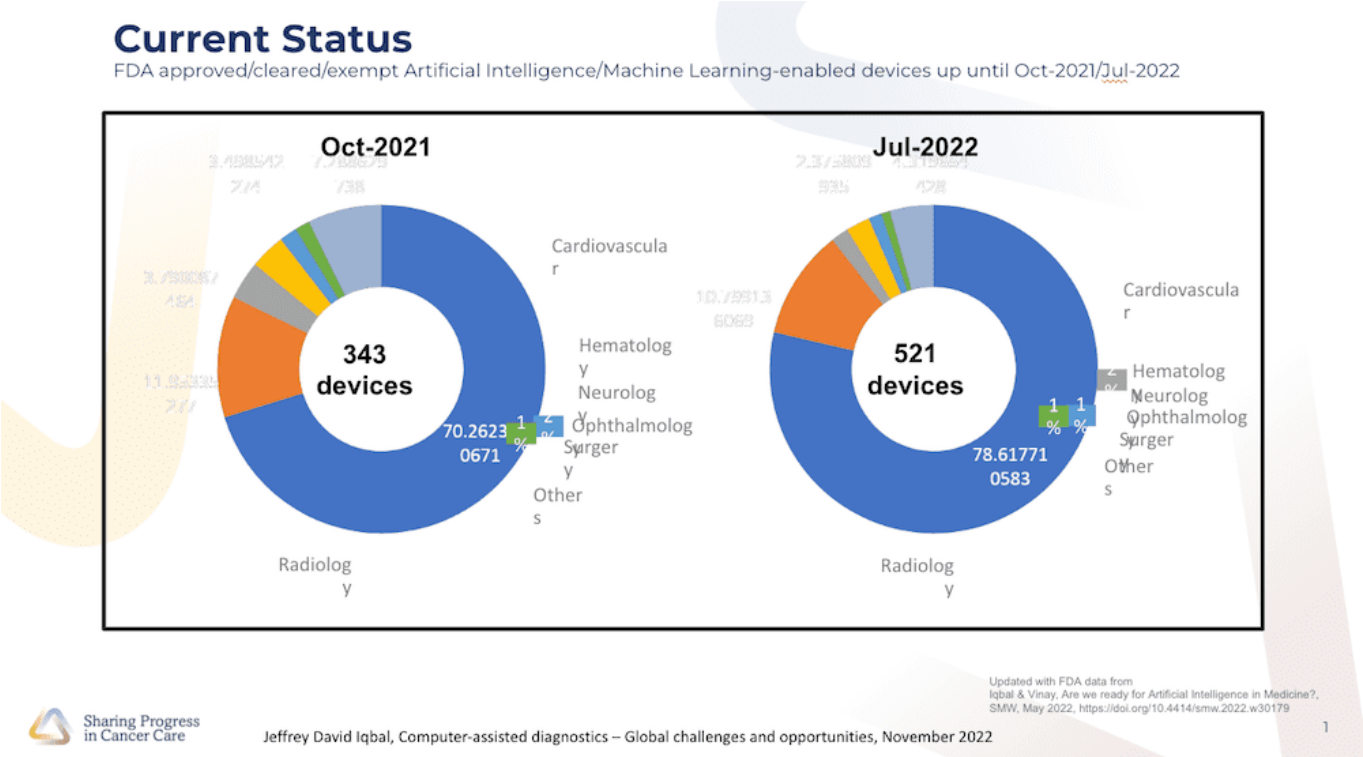


The second webinar in the SPCC Artificial Intelligence in Cancer Care 2022 series took place on 23rd November 2022. It was chaired by **Viktor Kölzer**, Attending Pathologist and Assistant Professor, Institute of Pathology and Molecular Pathology, University Hospital and University of Zurich. This seminar is held at an exciting phase in medical diagnostics and in medicine in general. We are witnessing a paradigm change where the digitalisation of medical diagnostics offers enormous opportunities for precision medicine. Doctors and engineers are coming together as joint drivers of innovation in this sector to create new methods for active diagnosis and precise treatment of patients, in cancer in particular, but also in other diseases. Where can this innovation lead us? These new data-driven techniques will be able to support precise and detailed analysis in clinical diagnostics. We have entered the era of big data where the possibility of digitising clinical, laboratory imaging data is becoming increasingly important and enables us to use clinical support tools driven by advances in machine learning and computer science. Digitalisation becomes a new tool, but also a connecting principle between medical specialties in future medical practice.

Computer-assisted Diagnostics: Global Challenges and Opportunities

Jeffrey David Iqbal, is a Postdoctoral Fellow at the Digital Society Initiative of the University of Zurich, in Switzerland, working on Artificial Intelligence in Medicine. His research focusses on the applications of AI within current and future clinical settings. Dr. Iqbal started with a terminology definition for the title of his presentation. Diagnosis cannot be a standalone process. By itself, it does not provide much value to the patient. When we talk about diagnosis, we need to see it in the context of treatment, alleviation, prediction, monitoring, and the whole wider process framework. As for “computer-assisted”, there is a lot of talk about AI, machine learning and digitalisation and the concepts are often used interchangeably. Digitalisation is a very wide term. In our context, it is essentially a process of using computer systems for the provision or support of healthcare delivery. “Artificial Intelligence” (AI) is placed within this field: computer systems able to perform non-

physical tasks normally requiring “Human Intelligence”. And here we have a problem already because there is not a lot of clarity on what Human Intelligence really is. Within the area of digitalisation, there are further terms: Machine learning and Deep learning. Electronic health records (EHR), for example, are not really part of what we call artificial intelligence. So, the term computer-assisted diagnostics is something more open. It is definitely digital, it may be AI, but not necessarily.



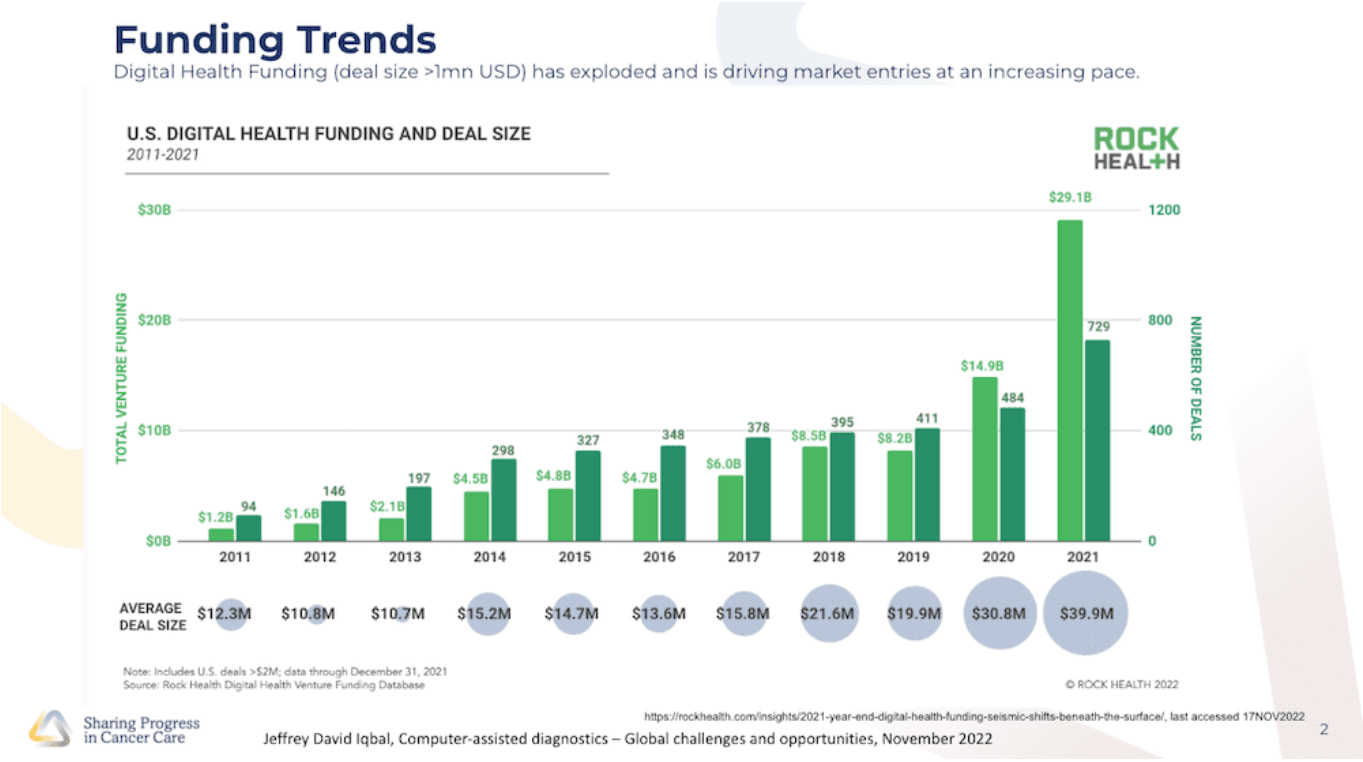
The vision of AI Diagnostics is to deliver greater value to key stakeholder groups: healthcare professionals, patients, and payers. For the healthcare professional, assistance in making a diagnosis can decrease the provisional burden. The patient, of course, can profit from better outcomes, and possibly a decreased interventional burden down the line, by having the right diagnosis from the very beginning. And payers may have a decreased financial burden, although there is no evidence yet that AI diagnostics can really bring costs down on a systemic level. But it is certainly one of its goals. The key underlying mechanisms of how this value can be provided are well known: a shift to earlier intervention, of time of pre-disease state versus success rate of potential intervention. We can also observe a shift along the self-care spectrum caused by digitalisation, including diagnosis. Another aspect of the value given by AI diagnostics is to enable the physician to cope with an exploding domain knowledge, at molecular, genetic, histological levels. Let’s think for instance of the advances made in the past ten years in our knowledge of DNA damage response (DDR). On a more histological level, non-small cell lung cancer used to be seen as one disease. Over time we have discovered more and more differentiation, it is almost like a syndrome, a deeply sub-grouped kind of disease. And this, of course, has implications for clinical practice. We are seeing an explosion of registered clinical studies. There are many different treatment protocols, sometimes even competing. It is very difficult for physicians and other healthcare providers to keep up with all of this, and computers, whether it is AI or not, can help us cope.

Another vision, perhaps 10-25 years down the line, is the hotly discussed **digital twin**. A digital twin is a live or near-live representation of physical entities, such as organs, body, potentially even the mind. It could help with the process of continuous screening, monitoring, and diagnosis in an automated way, moving away from the resource-constricted healthcare provision models we are in right now. The physical twin, i.e., the human, would constantly feed data into their digital twin

model. Events, actions would be recorded, for instance a surgical procedure, and there would be a feedback mechanism. The digital twin is not just about diagnostics and screening, it also shows how that individual acts.

The journey of computer-assisted diagnostics began back in the 1950s, and some of the articles written at that time even anticipated those ethical, legal, and societal issues that we are discussing today. A paper published on The Lancet in 1954 introduced a device capable of producing a list of potential diagnoses from registered symptoms. In the early 1970s, diagnosis implemented with an electrical computer surpassed human accuracy. DIALOG and INTERNIST-I, systems for computer assisted medical diagnosis, were developed, and then extended to covering general internal medicine. Between the 1980s and 2000s we saw a lot of activity aided by increased computing power and access, as well as algorithm development. But the real **game-changer moment** happened in 2009 with the American Recovery and Reinvestment Act (ARRA) that established incentives for Medicare providers to make "meaningful use" of EHR technologies. This led to a massive surge in data availability and usability in the US, but the rest of the world is following suit.

The FDA's list of AI/ML-enabled devices, up until October 2021, comprised of 343 devices. From then until the end of July 2022, nearly 200 more devices were added. Diagnostics, in the form of radiology, is at the forefront, accounting for 79% of those devices. Digital health funding surged over the last few years, peaking in 2021. We've seen a small slowdown in the first two quarters of this year, of course dictated by major economic factors, but the long-term projections are still very positive. As to the funding trends, the top funded value proposition is R&D, while oncology is among the six top funded clinical indications.



The **barriers to more widespread adoption** are to be found in the Evidence-Based Practice (EBP) lifecycle. We start with research, where we gather knowledge, then we enter the development phase, we come up with a device, drug, or procedure, and finally, we need to transfer it into clinical practice to reach the clinical endpoint. However, there is a research-to-practice gap. Less than 50% of those evidence-based practices make it to clinical practice and it takes an average of 15 years. There is also a regulatory gap, the lack of an effective regulation of digital health technologies. The FDA, but also other national players, are varying their requirements and of course, digital health

devices are very different from what regulators are used to, but also physicians and providers. The challenges in this field are not just on the technology side, the algorithms, the data collection, but also regulatory and revenue models. In sum, digital health technology is a vast category. Diagnostics must be seen not just as an isolated function, but in a more integrated way. We're seeing some key trends and mechanisms driving potential value, albeit still unproven at a systemic level. The digital twin concept may be a game changer. Computer-assisted diagnostics is not new. It goes back to mechanical computing devices of the 1950s and huge funding is still pouring in. With funding, we shall probably see some changes in the market, and new products being driven. The real question ultimately, remains adoption.

Digital pathology: Image analysis and machine learning in cancer diagnostics

Andrew Janowczyk is Assistant Professor, Biomedical Engineering Department, Emory University, Georgia, US, and Senior Research Scientist, Precision Oncology Center, Lausanne University Hospital, Switzerland and a Data Analyst at Geneva University Hospital. **Computer-aided diagnostics** (CAD) uses algorithms to help clinicians analyse data, such as fMRI, histology, X-ray, CT, etc. **Digital Pathology** is a transition from an analogue to a digital process. Previously, pathologists would look at tissue on a glass slide; that slide is being digitised and pathologists are now looking at those images on computer screens. Prof. Janowczyk showed an example of a typical output. Hematoxylin and Eosin stains (H&E) were used: H stains cell nuclei in blue, and E stains extracellular matrix and cytoplasm in pink. The 40x magnification image, the most commonly used clinical magnification, was 100k-by-100k pixels, resulting in a compressed file of about two gigabytes. If this image were to be uncompressed, it would coincide with roughly 32 gigabytes in raw data. These images are massive and potentially contain a wealth of information. **Why is digital pathology useful?** Almost all patients have an H&E slide in order to confirm their cancer diagnosis, and as a result there are vast amounts of these glass slides already in existence within the hospital's slide repositories and biobanks, most of which are kept for about 10 years for legal reasons. Digitising them is relatively inexpensive, typically costing only a few dollars per slide. More digital data is now being routinely created every day at an increasing rate. As more hospitals are purchasing digital slide scanners, they are scanning more slides, essentially building our cohorts for the future. Digital pathology improves the efficiency and robustness of medical diagnoses. It is fast and reproducible. **How can we use this data?** We can perform data mining to identify trends; identify subtle image patterns that may not be visually discernible; build systems to aid – not replace – pathologists through decision support. In fact, we will probably need more pathologists, to deal with the amount of data we are going to produce. The key concept here is the phenotype: the way in which the tissue presents itself is the total sum of all the underlying changes. Even if we do not know the RNA expression levels, even if there are methylation statuses that we do not know, if there are mechanisms that we are not even aware of, as long as they have some impact on the way that the tissue presents itself and in the way the body responds, we can now quantify and use that to predict therapy response or the prognosis of that patient. In the long-term, ideally, this will lead to a concept of precision medicine where we treat each person individually in the most optimal way possible based on large amounts of retrospective data that essentially finds a digital clone, instead of a digital twin, i.e., someone who is very similar to them and has previously been successfully treated.

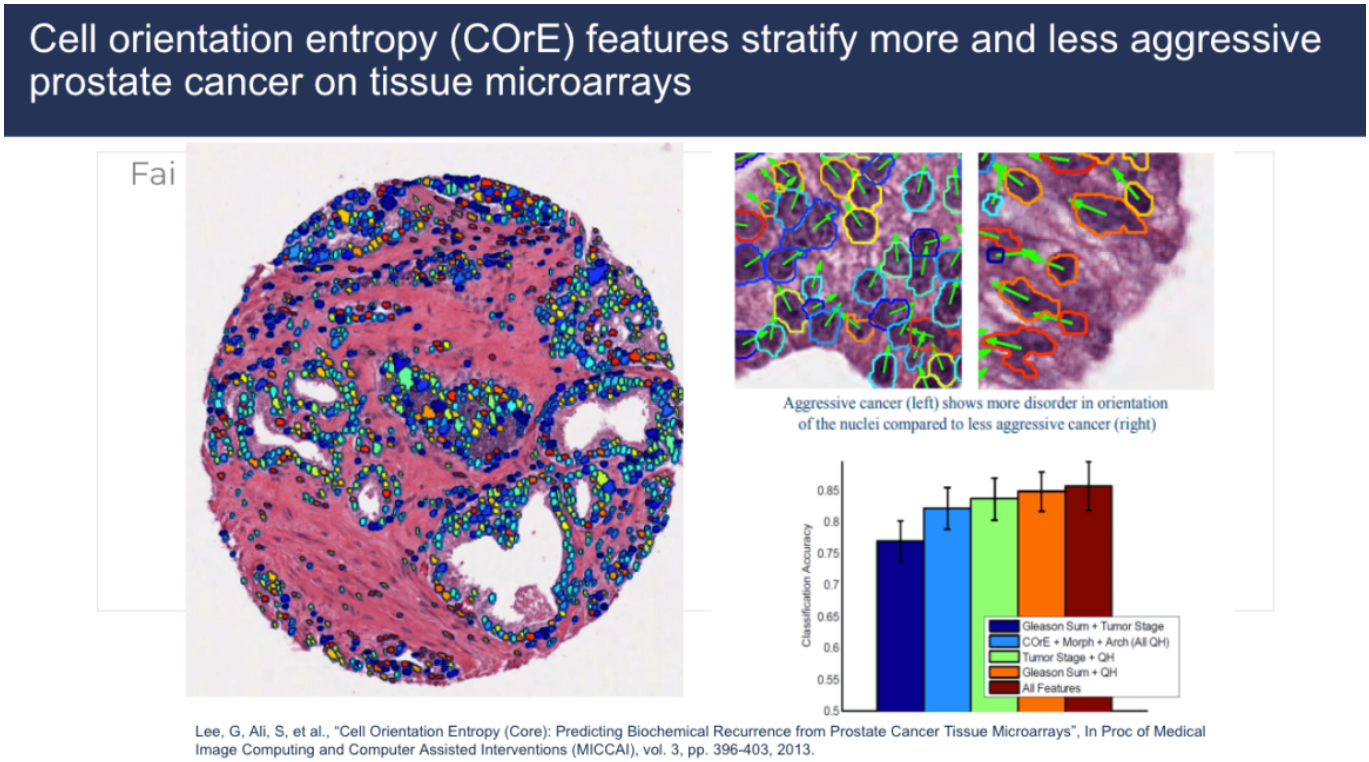
There is a **difference between research and clinical CAD applications**. In clinical applications, we essentially recapitulate and automate existing processes, tasks like cancer detection, grading, counting, area estimation, that are very laborious and subject to interobserver variability. We can use high throughput screening, whereby, instead of having to look through numerous slides to confirm our diagnosis, we can have an algorithm that finds the most relevant one first. CAD applications can bring improvements through quantification, reproducibility, and definition refinement. For example, cribriform is a pattern in prostate cancer that seems to be associated with

biochemical recurrence. The problem is that visually it is difficult for pathologists to agree on what it looks like, and that depends very much on whom they were trained by. However, we can have an algorithm identify this pattern. Professor Janowczyk's team in Cleveland did just that. They took a series of pathologists' annotations; they trained a deep learning algorithm to find this cribriform pattern, and what they saw was that they could actually start to stratify these patients by the likelihood of their biochemical response. This works best in the Gleason grade group 2, where patients are either going to be given active surveillance or more aggressive treatments. So, here we can even start to interstratify that group and provide better care for those patients.

On the other hand, some research applications are trying to extend beyond our current knowledge. They are looking at new features and metrics, striving to identify different novel subtypes, to connect with biological elucidation, and we expect to see improved patient care through augmentations and new insights. There is realistically a 5 to 10-year difference between these two categories of tools we are building. Another interesting example concerns cell orientation entropy (COrE). Researchers in Prof. Janowczyk's team found that when all the cells or the cell nuclei are facing in the same direction, this is associated with a less aggressive form of prostate cancer. While if the cells are facing in different directions or there is more entropy in their orientation, this is associated with a more aggressive form. This is a great example for two reasons. The first is that once you are aware of this pattern, which was discovered by an algorithm, not a human, you can go and observe it personally in a multitude of cases. This is something discovered by a computer that we can visually verify. Of course, a human could not look at a million cells on a whole slide image, estimate their orientation, and then extrapolate this to an entropy feature. But a computer can do so, very quickly and accurately. We can take this cell orientation feature further. We can take things like nuclear features, their size and shape, and they should have added value above common, let's say, tumour node metastasis grading schemes, and actually provide better stratification in terms of prognosis for patients. For instance, we looked at nuclear shape and orientation features from H&E images in early-stage Estrogen Receptor Positive (ER+) breast cancers to predict survival. This is the most common type of breast cancer in the US and identifying which patients will receive added benefit from adjuvant chemotherapy is crucial. Along similar lines, spatial arrangement of tumour infiltrating lymphocytes (TILs) and local density variance shows high correlation to the patient response to Nivolumab in non-small cell lung cancer (NSCLC).

What are the added values of digital pathology? On the clinical side, we can expect to have near-term deployment of many of these tools. H&E slides are already routinely being created every day all over the world for cancer patients, so, once the infrastructure is in place, we just need to have them scanned. Once they are scanned, the infrastructure can operate upon them and deliver a result. There is an opportunity here for large-scale cross-site validation of prognostic/predictive work, because we are creating large amounts of data, but we are also able to share that data fairly easily. It is worth noting that retrieving H&E slides from the archives and scanning them is very low cost. Typically, in the order of $1 or 2. This is quite different than trying to do some type of genetic sequencing. Digital pathology is inexpensive because we are only paying for computers, and the infrastructure is highly reusable. It is non-destructive: we can keep re-examining the same slide multiple times from different angles as the technology improves. It also allows for easier sharing of data. We can transmit digital slides over the internet to our colleagues, who will receive them presently. Interestingly, ethics requirements for digital pathology tend to be less strict than genetic information because the data is essentially anonymised by default. On the research side, we can now start to use digital pathology to examine large amounts of slides to help account for tumour heterogeneity. Costs in generating additional slides are minor. If we make 5 or 100 slides for a patient, the real expense is in the pathologist's time to evaluate those slides, not in the generation of the data itself. We can now use digital pathology to find representative regions or patients for higher-order approaches. So, instead of looking at 100 slides from 100 patients, we could start with

5 patients and a small region. We can do laser micro-dissection or spatial transcriptomics and reduce a lot of the noise that may be in our dataset. Digital pathology also allows for hypothesis generation and data screening. We can start to predict the genotype of patients directly from their H&E. In order to help facilitate this research, Prof. Janowczyk's group built a number of open-source tools – the **HistoTool suite** (e.g., histoqc.com, quickannotator.com, patchsorter.com, cohortfinder.com).



There are two parts in the "magic" of developing a biomarker from slides: a pre-analytical component and a post-analytical component. The post-analytical component is essentially user specific, and outside the scope of HistoTool. On the other hand, pre-analytical components are shared in every biomarker project. HistoTool provides an open-source preanalytical pipeline. The first tool in the kit is **HistoQC**, which allows for highly reproducible quality control, and has been shown to improve concordance between readers from about 70% to about 96%. **HistoBlur** allows for rapid and precise detection of blurry objects at scale, so we avoid performing computations on regions that are unlikely to work well with our classifiers. With **QuickAnnotator** we can annotate objects very rapidly, in some cases 100 times faster. **PatchSorter** is a rapid image labelling tool, shown to improve efficiency by 400%. All three of these tools are aided by deep learning, so they become agnostic to the types of stain and the tissue type, and actually can work even outside of cancer, for instance in kidney diseases, or even in predicting heart rejection. Lastly, we have **CohortFinder**, a tool for intelligent data partitioning using quality control metrics.

Touching upon the **challenges**: there is still a lot of **inter-site data variability**, such as staining variability, scanning differences, protocol differences. There are **infrastructure limitations**, like storage volumes, integration into clinical routines, access to clinical metadata. And, of course, we would like explainability, we would like to be able to understand what the features we find actually mean.

Digital radiology: AI guided cancer diagnosis and treatment planning

Bettina Baeßler is Professor and Head of Cardiovascular Imaging and Artificial Intelligence at University Hospital, Würzburg, Germany. Radiology abandoned film and went fully digital very early

on, so, compared to other medical disciplines, it is quite ahead in AI development. However, there are still many potential applications for AI in radiology, not only in diagnostic decision making or image detection segmentation, but all along the entire imaging workflow, beginning with the indication and scheduling for a radiology exam, image acquisition, and reconstruction. For example, MR Imaging needs reconstruction of the images, and this could be AI-assisted today. Segmentation, detection, and quantification as well as diagnostic decision making could be aided by AI, but we can also go even further and have AI-assisted reporting and communication, as well as prognostic assessment. So, there is quite a wide range of potential applications. Not all of them are currently addressed. Although the main focus is on the analytical and diagnostic part, the workflow aspects at the beginning of the pipeline are very interesting, and hopefully more applications will be coming out for speeding up this process and eliminating the hurdles we tend to see in the clinical setting.

In oncological imaging there are currently three main AI applications. There is **detection**, i.e., detecting regions in the images which are not normal. There is **characterization**, subdivided in diagnosis (defining abnormalities as benign or malignant); staging (assigning patients to particular categories) and imaging genomics (linking imaging to genetic and molecular features). Finally, we can use AI for assessing **treatment response** beyond what is normally used in clinical routine, such as, for example, the RECIST criteria (Response Evaluation Criteria in Solid Tumours), which are known to have some limitations. All of these different kinds of data sources and medical images are set to enable modern healthcare strategies. When it comes to precision medicine, all of these data are very relevant, and medical imaging is only one part of the whole, but a very interesting part.

There are many different omics clusters: genomics, epigenomics, transcriptomics, proteomics, and so on. Oncological radiology has also an omics cluster, called **radiomics**. This is a method that uses information about the spatial distribution of Gy values in medical imaging to extract mathematical features, quantitative features, biomarkers, feed them into the big data pipeline, and then mine them for improved decision making. Our current criterion for assessing oncological response is RECIST, but unfortunately important changes stay undetected with this method, and this is where radiomics comes in, allowing us to move to more deep learning-based approaches. We can have our images, but we can also annotate them, not only by measuring size but also by drawing regions of interest in our images manually or automatically. We can do different things with this. One is radiomics, another is deep learning, and there is also something called **habitat imaging**, which looks at the micro-structure of, for example, an oncological image. We know that tumours are heterogeneous, and often have different clonal compositions which might also change during treatment. The idea was that maybe our images can reflect this intra-tumour heterogeneity because we can see there are different areas in the image, and by extracting these mathematical numbers, we can also display those features as a map. The idea was to get more information out of the image, which reflects, for example, histopathology or clonal status. How does this work? We take an image, for example, an MRI or a CT scan. We do a segmentation which can be done in different versions: 2D, 3D, manual, semiautomatic, fully automatic. And then, we extract those features by mathematical procedures of varied complexity. There are four main types of radiomic features: morphological (shape and volume); statistical, which is further classified into first-order statistical features (histogram) and higher-order statistical features (texture); regional (which represent intra-tumour clonal heterogeneity through subregional clustering); and model-based (extracted using mathematical approaches). Then, we can apply additional filtering on the images, for example, enhance edges. Professor Baeßler took an example of a radiomics approach for differentiation of benign versus malignant lung nodules, which showed how much more information we can glean compared to the CT alone. We can also predict progression-free survival. Some radiomics features have been shown to be better predictors for progression-free survival than the standard features of age or gender or performance, or even clinical performance status. Furthermore, we can predict mutational status. Still looking at lung cancer, there are features associated with, for example, EGFR receptor

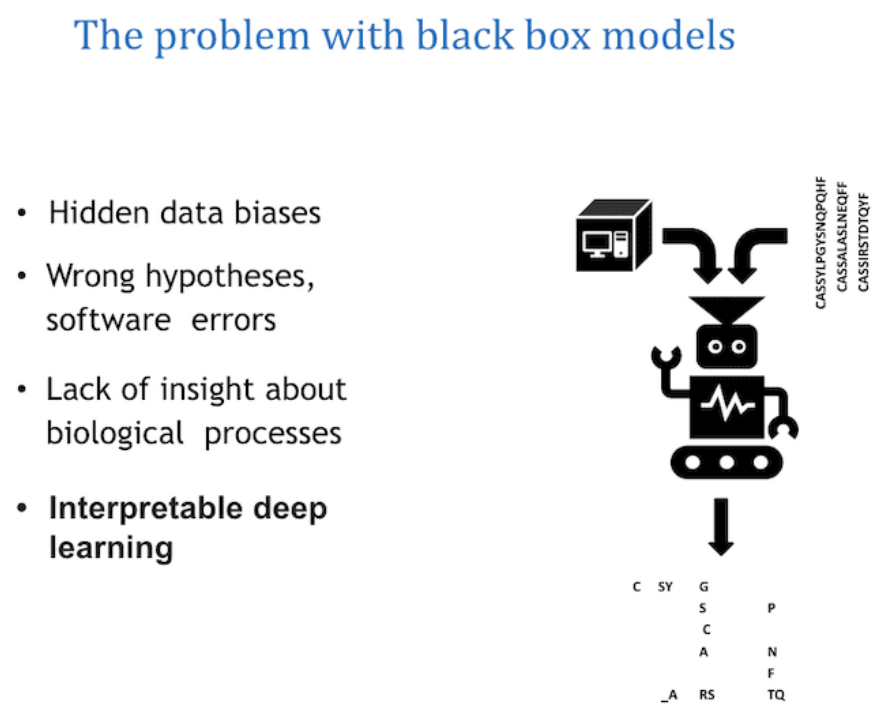
positivity or negativity. So, there are many correlations between radiological features and genomic or histopathological ones. If we do not manually extract those features but put the images directly into a pipeline, then, we can do deep learning. There is a wealth of studies for deep learning in medical imaging. One field where DL is used a lot is lung cancer detection. An algorithm is trained to detect candidate nodules, which are then classified in benign or malignant tumours. Other fields where DL is often applied are breast cancer detection, and rectal cancer, where the segmentation is combined with a tumour or non-tumour probability map. There are a number of FDA-approved software available, mainly for breast lesion characteristics, pulmonary nodule detection and prostate lesion characterization.

Of course, there are also limitations and challenges, one being standardisation. Many technical factors influence the extracted radiomic features, so, radiomic studies often cannot be applied to a different dataset. Another big challenge is data anonymisation. A very interesting study was conducted, which was also highlighted in the Wall Street Journal. Researchers managed to identify the patients from MR images of their head by reconstructing their features with face-recognition software. This is a huge challenge that needs to be addressed. For better data protection, we must look at learning infrastructures. With **local learning**, all the training is done locally and is not automatically shared with other institutions. This is obviously very limiting. In **central learning**, data and algorithms are sent to a dedicated server, to which all the hospitals have access, and the information is shared. This structure is impractical both in terms of data increase and data privacy. It is not secure, and for that reason it is not allowed in the European Union. **Federated learning** was until recently the main solution. The systems are trained locally, and the raw data stays locally. Only the weights are transferred to a central location, thus resolving the privacy issue. The learnings are just numbers and do not reveal information about the patient. The final model is then combined using these weights. The **RACOON project** in Germany is an example of such a federated learning infrastructure: incentivised by the Covid pandemic all the German university hospitals initially contributed their data on lung disease but are now extending the dataset. We can even go beyond federated learning into **swarm learning**, which eliminates the central server and uses blockchain technology, so it is even more secure. We are currently combining a swarm learning approach with generative learning diffusion models to get synthetic data which we can then share with, for example, the industry, for training reliable models. This is quite an ambitious project, but it is a very interesting new approach in medical imaging.

Interpretable deep learning in biology

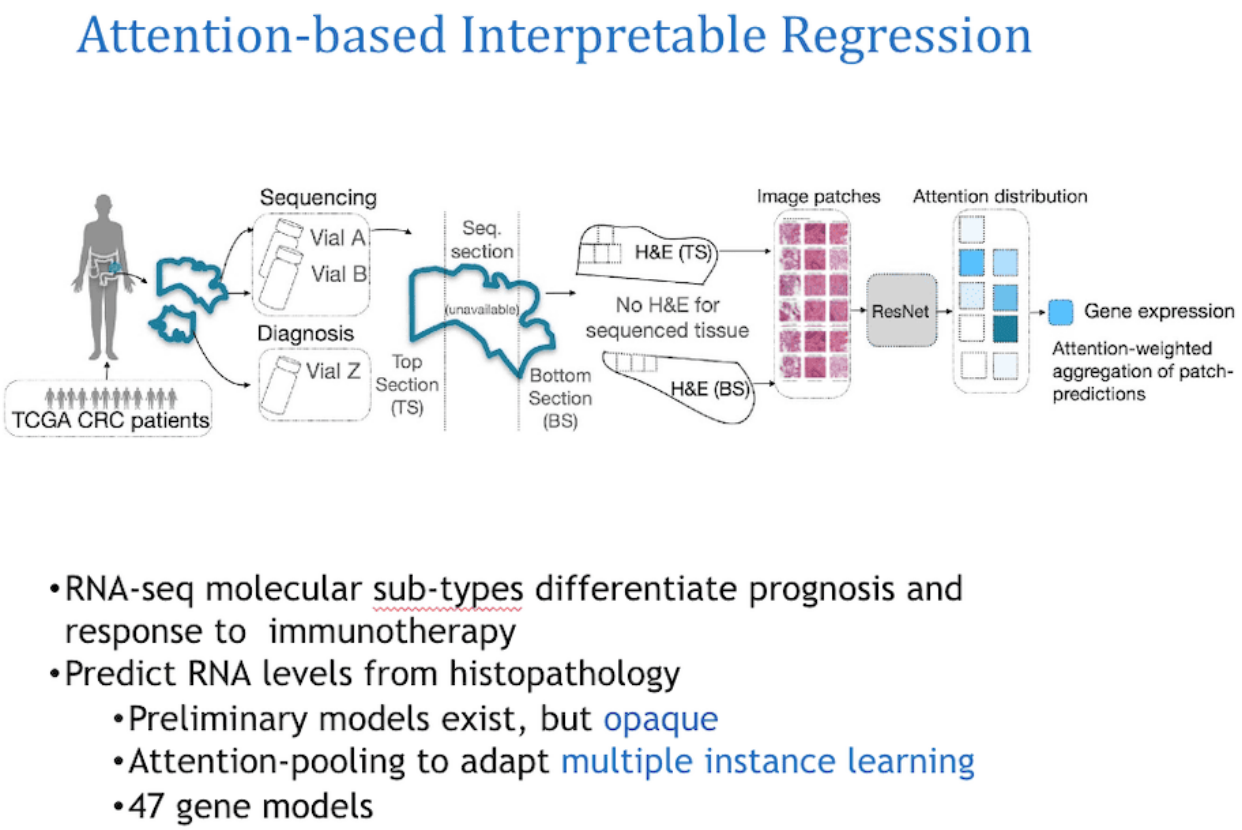
The final speaker was **María Rodríguez Martínez**, Staff Member at IBM Research, Zurich. One of the main recent achievements for biology probably is **AlphaFold**, an AI system that can predict protein folding with astonishing levels of accuracy. Unfortunately, DL models behave like black boxes, and this causes problems. For instance, hidden data biases. We may have cohorts that are extremely skewed, perhaps gender or ethnic biased. Wrong hypotheses may be formed, we expect some treatment to have a good response based on other cohorts, but it actually does not work in a different population. There can also be hidden software errors, quite often the model gives you a prediction that seems reasonable but is actually wrong. Even if the model were perfectly good and accurate, with unbiased data, scientists would still want to understand the mechanisms. If we are stratifying cancer patients, we want to understand, for instance, why the model thinks that a given person should receive aggressive therapy. While we may not be able to make deep learning models fully transparent, there is a midway approach, called **interpretable deep learning**. IBM has been developing this approach for different applications. For instance, **PaccMann**, a DL model to predict drug sensitivity on cell lines. What is important about this model, in this context, is the interpretable aspect. The way we achieve it here is through **attention mechanisms**. This is something that you include in your model when you are building it. And then, the model gives higher weight to those

features that are important to make a prediction. Let's take for example two different compounds, masitinib and imatinib. Both of them are FDA approved for the treatment of leukaemia. The two compounds are almost identical, except for one atom. In one you have sulphur and in the other nitrogen. In masitinib we see that PaccMann is quite certain that this atom is important, most of the attention weight is concentrated on it. While in the case of imatinib, we see that the attention is here and there, with no winning pattern. This is qualitative. It is still not a full explanation of how these compounds work, but by comparing the two, we can at least draw a new hypothesis.



New techniques have enabled sequencing of T-cell receptors (TCRs) and their antigenic targets (epitopes), and we can now research the missing link between TCR sequence and epitope binding specificity. **TITAN** (Tcr epITope bimodal Attention Networks) is a **multimodal neural network** that encodes both TCR sequences and epitopes separately. We can then ask the model to give us the binding likelihood. But we can also extract insights from the attention mechanisms. We can get an idea, for example, of which amino acids may be important for the binding. Now, the other important aspect of interpretability is that it can also tell you when things do not work as expected. This was a harder test, because of lack of data. But the interesting thing was that the attention maps clearly showed that something was not right. Interpretability can tell you when things work well and help you elaborate new hypotheses, but it can also tell you when things do not work, and in science, this is equally important. Going back to more image-based digital pathology, IBM has developed interpretable models to predict the focus on colorectal cancer. Often, score grading is based on different RNA-sequencing agents, but usually the most abundant data type for patients is images. Can we try to predict gene expression values from images? Yes, it has been done, but the accuracy is not ideal, and the models are opaque, we have no understanding of how they work. Again, interpretability can improve, at least somewhat, the prediction. In a study, Dr. Martínez's group used the Cancer Genome Atlas (TCGA) data. Each image is divided in different patches. A technique was used, called **multiple-instance learning**, in which you assume there is a signature somewhere, but maybe not all patches are going to have it. Some of the patches in a whole image slide might be perfectly normal, the cancer signature might be only associated to some of the sub-patches. With multiple-instance learning we use attention mechanisms to see the gene expression level of each sub-patch, and then use that attention score as a sub-weight. We give higher weight to the patches that the model finds more informative. To quantify the accuracy, for each patient we get a predicted

value and then, we can compare them with the measured value because we have the RNA-seq data from the CGA. And then, instead of just having the absolute distance, we can compute percentages. This is also comparable across patients and models.



With interpretability we can enhance prediction and learn from error. These models have proven to perform 30% better than current state-of-the-art models. Of course, what we are doing is extremely hard. We are asking the model to predict gene expression values from an image, and we cannot expect perfection, but the ability to visualise attention scores gives us rich information about what the model finds important. In conclusion, deep learning is achieving breakthrough performances, but models are black boxes, we have no idea what is driving the prediction, thus we fail to gain insights about cancer, why this patient is stratified or why this therapy is working on this patient. Also, very importantly, these black boxes might be hiding data or algorithmic errors. Interpretability allows us to extract new insights about biological mechanisms and helps identify algorithmic/data biases. Interpretability improves prediction accuracy and generation of visual hypotheses.