Subtitles and transcriptions

Subtitles and transcriptions are available for selected materials for purpose of helping users understand the contents of the educational sessions.

Uncertain words have been indicated with ?? before and after the part. Parts that could not be understood at all have been indicated as [Audio Not Clear].

Every effort has been made to faithfully reproduce the audio of the sessions as recorded. However, no responsibility is accepted for mistakes or omissions. ESO does not endorse any opinions expressed in the presentations.

Evaluation of kidney lesions using AI tools

Prof Balis: Hi, my name is Ul Balis. I'm a pathologist at the University of Michigan. I'd like to thank the organizers for the opportunity today to share with you some recent innovations in digital pathology as it applies to computational image analysis in nephropathology. Recognizing that these presentations are short, I will focus particularly on one exemplar, which is the NIDDK Kidney Precision Medicine Project, as a vehicle by which we're using some of these newer tools to accelerate both discovery and classification of whole slide imaging. So as background, the NIDDK project, the Kidney Precision Medicine Project, is unique, in that it is both an investigative study, but also an attempt in terms of clinical management, by having actual patients, not just research participants, who are donating research biopsies that can be followed longitudinally to better understand two specific classes of kidney disease, acute kidney injury and chronic kidney disease. And one of the major aspects of this effort is creating a Kidney Tissue Atlas, and in tandem with that, finding current and new disease subgroups, as indicated by both the morphology from digital computational analysis of the whole slide images, enriched, however, by additional modalities of data. So, this is a multiplex study, and I'll expand on that in a minute. So, this is a ten-year timeline. We're currently practically finishing the first five-year period, so that has allowed us to establish the Kidney Atlas itself, and has allowed us, also, to bring on board a number of tissue interrogation sites that collaborate with a central hub in providing next-generation analysis techniques which are multiaxial. That is to say, they complement the histology itself. However, it should be noted that histology is a central anchoring point for this study in allowing us to better characterize the various kidney diseases. So if we think about the atlas itself from a conventional point of view, one would think of, for example, a book with collated photographs of the various normal and diseased states of the organ systems, in this case, kidney, that becomes, in the more modern setting, a digital atlas of whole slide images, but this is now enriched with multiple classes of metadata from within the images themselves, as well as outcomes data and comorbidities like matched cases, matched both by the images themselves, so-called content-based image retrieval, and matched also by metadata associated with images, and also with the other types of multiplex molecular content. And this now becomes a vehicle by which, when you interrogate the image information, for example, a diseased glomerulus, you can now match that with pure cases, match that, also, with survival information and with molecular data, generating a cohort which is more richly informed, not by merely the single case itself, or one modality such as imaging or Omics. but the Omics, the imaging, and the outcomes and clinical course data all in concert, which allows for a much richer and well-matched cohort extraction of similar cases, which has, potentially, significantly better discovery and statistical power for analyzing a disease. In terms of a foundational model of image analysis, obviously, this is a pyramid that has at its foundational level just the operational task of having whole slide

imaging workflow in place. We have that now, through the consortium, so that we can generate images from the multiple slides that are arriving from multiple sites generating these biopsies. The quality detection systems are now in place. We have made use of the HistoQC application, which is freely available from Dr. Anant Madabhushi's lab in Case Western. That tool allows for detection of all the common spurious types of artifacts within histology, and that allows for the earliest possible correction, whether it be re-scanning or even generating a new slide section. And then in terms of a foundational image analysis, the first and most important ability is to identify individual features, but then combine them to create a multivariate analysis of the image itself, and one example tool I'll be showing you a little bit later in the presentation VIPR, allows for the analysis of the entire surface area of images at the pixel level and generating interactively with a pathologist or life scientist driving the application a computational pipeline, which results in a generalizable tool for segmenting and then characterizing the entire surface area. And this is important, as I'll mention in a moment, simply because the effort of segmenting images, which is a needed step in training of machine learning algorithms, at least can be extremely time-consuming in the setting of a subject matter expert expending manual effort and time to manually label large surface areas. The fact is, you can use that expertise to create exemplars by which, then, an algorithm will then amplify what a single individual could do across an entire whole slide image, or in fact, an entire entire library of whole slide images. So, the opportunity here is to take that expertise and then amplify it computationally. However, in terms of the so-called holy grail of advanced image analysis, which is a turnkey system which yields all the diagnostic recommendations, this is not what I'm speaking of, and I don't know that this will actually happen in my lifetime or in the remote future. I think the practice of pathology will still always benefit from expertise, using these tools in an informative way. I've heard it described as making sure that there's always an adult in the room. So, I don't suggest at all that these tools will be replacing the practice of pathology. They simply will change how pathologists practice, giving us more powerful tools, especially in the setting of multi-variate analysis. This is a motivation for why we need these high-performance, high-throughput tools. If you look at the practice of pathology and lab medicine from the prism of increasing levels of complexity of the data we generate, some sobering realities. Just simple chemistry values, chemistry panels, generate an initial cognitive threshold of between seven and ten data elements. Anything above that really needs computational support to extract the maximum information. And another sobering threshold which involves libraries of whole slide images and NGS data sets, this, when you hit somewhere between 10 to the 10, 10 to the 11 data elements and beyond, you really are into the realm of supercomputing and the need for high-performance computation, the tools, the storage, the stewardship model, and yet, as a practice of pathology, I would offer that we're far behind in training our cohort of pathologists and associated life scientists to be effective at handling HEDIS scale datasets. So, this is just a parenthetical observation that we have lots to do in terms of our data stewardship activities, and because of this, we clearly need simplified tools that handle large-scale data, and that will be one of the examples I'll show you very shortly. So just as a very, very brief refresher on supervised and unsupervised learning, in the mode of supervised learning, this is an opportunity for the subject matter experts to provide ground truth maps or exemplars to an algorithm so then an algorithm can then better apply those rules at a larger scale. As opposed to that, the unsupervised model simply looks for the naive and innate clustering of data, K-means being an example, and then presents those clusters to individuals, potentially identifying unusual, rare diseases. So, the example I'll be showing today is in the class of supervised learning, recognizing that it is intractable, typically, for a pathologist or a line of scientists to spend the needed time to segment entire libraries of whole slide images. And also, within machine learning, in terms of creating models, this is, of course, the standard model, where once you engineer the features and train the initial model, you look how it operates in the real world, you deploy it, and then based on its failures and successes, you update the model and reiterate, constantly making a better and better model. And that is exactly how the tools that are being developed for the Kidney Precision Medicine Project are being developed. As this very, very simple example, and this is from colon, obviously, not renal, the challenge is simply a case of training an algorithm to identify colonic crypts, a simple exercise that could take many, many hours to segment, have a subject matter expert literally circle every feature of interest. This could similarly

be tubules within a renal biopsy, and you can even see that even doing so, there's imprecision in what a human can do. What is really needed is segmentation at the pixel level across the entire surface area of the whole slide image to support high-throughput computation. I should just mention parenthetically that if your ground truth maps are imprecise, when you apply those ground truth maps to any potential pipeline, the imprecision you have in the results in terms of a region of interest mapping to a area under the curve, for an ROC curve, could be just as much a result of the imprecise ground truth maps as it could be of the algorithm failing, and then you won't know if your algorithm is good or bad because of the bad ground truth maps or because of the algorithm itself. There are two tools that are associated with the Kidney Precision Medicine Project for rapid segmentation generation. One comes also from the Madabhushi group. This is the Quick Annotator, which was recently published, in November, actually just a few days ago, and additionally, this tool is designed, basically, to take a small number of examples and then identify in terms of a patch setting, similar features that match the statistical criteria of the exemplar patches. And it's meant to operate in a fully autonomous fashion. Similarly, with the use of the tool that we've built at the University of Michigan, VIPR, for vectorized identification of prescreened regions, it operates in a similar fashion. It has an integrated high-performance whole slide image viewer, and it allows for convergence on a ground truth map generation with far fewer fields than you typically would need with convolutional neural networks. This is really quite important, because in the case of rare entities or unusual lesions within histopathology, you may not have the benefit of having thousands of images, and in that setting, it's ideal if you have an algorithm that will converge with high efficiency with the fewest number of examples or fewest number of fields of view, and this algorithm satisfies that criteria. Similarly, the algorithm has built-in integrated analytics for taking the individual features and then clustering them in a dimensional space using well-understood Python libraries such as k-means or XGBoost, plus other libraries, so that you can very quickly arrive upon a generalizable computational pipeline for converting a whole slide image into a map where every pixel location is no longer a color information, typically, it would be, otherwise, RGB, but rather is the biological context of that pixel location. So, it essentially is a pipeline for contextualizing image morphology at a pixel level into various predefined morphologic meanings, which is perfect for the downstream tasks of other machine learning algorithms that would benefit from knowing what each pixel represents in terms of biological structure. And this can operate at various length scales, the organ structure cells, or subcellular, and the algorithm in this demonstration in the next few slides will underscore some of that opportunity. So in terms of what this workflow looks like, we want to, again, contextualize pixels from being colors to biological context, so we start with an original image, here is a high-magnification image of renal tubules, for example, some normal, some vasculature associated with it, and a tangential cut of a glomerulus, just as a brief example, and then we generate multiple individual vectors which are enhancing for one or more particular features. Any one of these vectors alone, however, is not diagnostic, they need to be combined by the XGBoost or K-means to generate your final contextual map. So, this single vector, then, gets enriched by multiple vectors, so this would be, for example, the lighter areas, recognizing a vector has found nuclei, basement membrane, capillaries, tubular cytoplasm. And you continue doing this till you have, essentially, a map of the entire surface area. That gets generated into a metaspace of analysis in the high dimensions, since each one of these represents a high-dimensional space in and of itself, but they can be combined. Once cluster analysis is carried out, each pixel is carried out with analysis, then, in this high-dimensional space, yielding a classified image. So, in this simple example, we've identified quite effectively tubular epithelium, the epithelium tuft, and the nuclei of the basement membrane in the white space, as well as the glomerular regions in the sangium. Again, for a human to do this at the pixel level really would be intractable. In terms of feature engineering, the way this is done is, a user, excuse me, interacts with a simple user interface, generating a small number of patches, again, similar to the Quick Annotator tool as already described, allowing for the algorithm to make a generalized pipeline. As a simple example, in an exercise carried on in the early stages of development with this application, this is effort carried out with Dr Laura Barisoni at Duke University, giving a small number, and this is, of course, for a production pipeline, you'd never use this few examples, eight positive and negative examples of subnucleolar vascularization of tubules, a sign, presumably, of injury

and repair, and after training, we were able to demonstrate with exquisite precision at the pixel level an algorithm that would find the subnucleolar vacuoles, and this was mapped against what was, in fact, a conventionally-trained ground truth map made by several expert pathologists, which took an extremely long period of time to validate, but the ROC curve for this was better than 0.9, which is very, very good for a very unique feature. And this is one example of the specific pipeline that's possible. In terms of showing in high magnification what these algorithms actually look like, this is a representative renal biopsy. When converted from the PAS stain to the computational stain, even at low magnification, you can see that there, quite a bit of precision's been carried out. This has been entirely carried out by computational analysis. In zooming in, it's worth pointing out the consistency with which such algorithms can annotate and label the entire surface area. This particular pipeline was enriched for obliterated Bowman's capsule, and you can identify the Bowman's capsule obliteration here. And interestingly, this tangential cut is not a tubule, this is a tangential cut of an adjacent glomerulus, you can see some of the glomerular tuft, and the obliterated Bowman's capsule. Again, so the algorithm is able to identify with exquisite precision, biological context, based on the characteristics of these vectors at high-dimensional spaces. This is the application itself. I'll briefly switch to the application to show how a pre-rendered pipeline can be fine-tuned for a particular structure. This is a live presentation, so I won't necessarily go through all the uses of the application, but it is a live whole slide image built in where you can change magnification. I'm going to take this field of view and carry out a pipeline initial analysis. The pre-made vectors are identifying individually. However, these individual vectors, again, aren't the final output, they're the individual contributors to the dimensional space. This is, for example, identifying glomerular [Audio Not Clear] that was a tuft of the tubules. This is the glomerular region, small glomerular region, additional basement membrane, epithelium. This was nuclei, this is basement membrane, this is cytoplasm of tubules. And the reason I'm being able to identify these is the lighter areas are the areas that are statistically mapping. Worth pointing out again that any one of these vectors sets is not pathognomonic of characterization, but when you have, at the end of this process, 500 or 1,000 variables, this is white space, by the way, those variables combine to create a very rich classification. So, at this point, we've mapped everything. However, the subject matter expert doesn't see the mapping yet until we can turn on the features individually and decide if they're contributory. So that, for example, is a white space, that is capillary loops, that's also luminal space, more luminal space. You can see that the algorithm has identified literally every pixel, and classified it, and it's really a very quick process for the subject matter expert, now, to annotate or select which features are of interest. So, I've just labeled, now, tubular cytoplasm, that's worth keeping. This is the vector set I enriched for the obliterated Bowman's capsule. I'll keep that. That's my basement membrane, I'll make that red. Did a good job. That is the rest of the cytoplasm for the tubule. That is the remaining portion of glomerulus. I think we have a nucleus coming up soon. That's actually capillary loop, I'll turn that off. That looks like it's the rest of the cytoplasm of the tubules. It is. And that is the final portion of the tubular cytoplasm. And that's the nuclei. I think the last two are gonna be the remainder of the Bowman's capsule. No, that's not. This must be it then. It is. I'll make that seven. So, with relatively little effort, we've made a pipeline, now, that can classify this biopsy or library of biopsies, with this level, pixel-level precision. Obviously, there's not enough time to go through the rest of the details, but I will offer that this tool, VIPR Studio, is available to any collaborating groups. This is not a commercial product, this was an academic endeavor, and we are hoping to share this with as many interested parties as possible. Switching back to my presentation, to sum things up, we've also had the opportunity to use these mapping pipelines to create educational tools, so if you know the biological context of every location in an image, you can create additional layers of metadata on the image, and when, for example, a student moves around within an image, the appropriate area on a cartoon can be highlighted, which creates fascinating possibilities for interactive educational maps. Additionally, this is work carried out with Michael Eden at the University of Indiana, there's opportunities now for taking the spatially mapped biological context where pixels become biological context and map them with spatial transcriptomics. So, in the Visium Technology, with Visium, this is now 40-micron, and soon there'll be available, there'll be much smaller micron pitch, it's possible to generate 2,000 or more transcriptomic signatures per location, which then turns into a unique signature. And the goal is really to

develop a set of vectors that can correlate with the spatial transcriptomics, so that once that initial discovery is carried out, the morphology can be a surrogate for the underlying transcriptomic data, and this can become a powerful tool for enriching the type of molecular information that histology is able to extract from images. Similarly, once metadata is available in terms of transcriptomic signatures and patterns with images themselves, it becomes possible to create a crosswalk by which the genes themselves can be mapped to biological context of pixels which is stored in this image in the form of metadata layers, so as that various genetic signatures are selected, the regions that they correlate statistically can be highlighted, and this potentially can be a very powerful tool for both discovery as well as for diagnosis. So, this is a tool that's in pilot prototype phase, but we expect that this will be one of the products of the Kidney Precision Medicine Project as well. And similarly, it's possible to invert this so that you can go from the spatial realm back to the biological realm in terms of genes, So here selecting areas of a moderately normal glomerulus, and then submitting them to an imputation engine for identification of the associated genes by matching the metadata in those stored areas, again, the type of tools that will be available very shortly as a result of mapping transcriptomic data, other types of omic data, in a multiplex fashion with the histology itself. So, in summary, I'd like to thank you for your attention, and recognize my collaborators at multiple universities, particularly Matthias Kretzler, Jonathan Himmelfarb, the PIs of the NIDDK, Kidney Precision Medicine Project. Thank you very much.