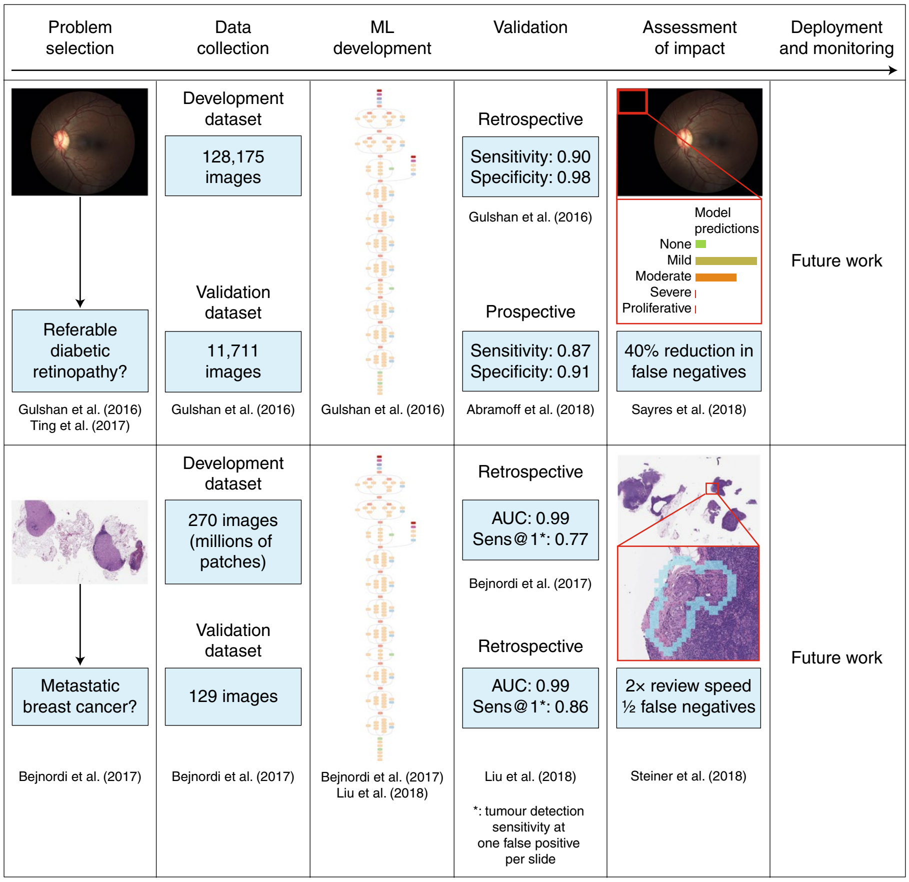
How to develop machine learning models for healthcare

Rapid progress in machine learning is enabling opportunities for improved clinical decision support. Importantly, however, developing, validating and implementing machine learning models for healthcare entail some particular considerations to increase the chances of eventually improving patient care.

Po-Hsuan Cameron Chen, Yun Liu and Lily Peng

Advances in machine learning (ML), faster processors and the availability of digitized healthcare data have contributed to a growing number of papers describing ML applications in healthcare. A common goal of these ML models is to improve patient care, both for clinicians and patients. For an in-depth discussion of how recent advances in ML work, we refer the reader to reviews[1].

In this piece, we will discuss the importance of the intended use of the ML model and its role throughout the process: problem selection, data collection, ML model development, validation, assessment of impact, deployment and monitoring (Fig. 1). We will focus our discussion on ML models for diagnosis (disease presence) or prognosis (risk of future outcome), both of which involve predicting a label based on input data. These principles are applicable to other clinical applications such as image segmentation for radiation therapy planning and measuring cardiac parameters from echocardiography. Applications such as drug discovery that are earlier in the bench-to-bedside journey will also share considerations in the data collection and ML model development and validation. However, these applications will have an increased emphasis on other types of validation such as experimental 'bench' work to validate computational hypotheses.



Problem selection	Data collection	ML development	Validation	Assessment of impact	Deployment and monitoring
Referable diabetic retinopathy? Gulshan et al. (2016) Ting et al. (2017)	Development dataset 128,175 images Validation dataset 11,711 images Gulshan et al. (2016)	Gulshan et al. (2016)	Retrospective Sensitivity: 0.90 Specificity: 0.98 Gulshan et al. (2016) Prospective Sensitivity: 0.87 Specificity: 0.91 Abramoff et al. (2018)	Model predictions None Mild Moderate Severe Proliferative 40% reduction in false negatives Sayres et al. (2018)	Future work
Metastatic breast cancer? Bejnordi et al. (2017)	Development dataset 270 images (millions of patches) Validation dataset 129 images Bejnordi et al. (2017)	Bejnordi et al. (2017) Liu et al. (2018)	Retrospective AUC: 0.99 Sens@1*: 0.77 Bejnordi et al. (2017) Retrospective AUC: 0.99 Sens@1*: 0.86 Liu et al. (2018) *: tumour detection sensitivity at one false positive per slide	2× review speed ½ false negatives Steiner et al. (2018)	Future work

Fig. 1 | Examples of different phases of the translational process of developing, validating and implementing ML models for healthcare. The first example shows studies for predicting referable diabetic retinopathy from retinal fundus photographs. The bottom row illustrates examples for the detection of metastatic breast cancer in lymph nodes in histopathology slides. ML development images adapted from ref. [31], Google. Other images created from the data in the indicated references.

Selecting the appropriate problem

The first step in developing ML models for healthcare is problem selection and defining the prediction task. Successful ML models should be expected to make a meaningful impact in patient care by providing actionable insights. For example, an ML model can be used to predict factors that are used in clinical treatment guidelines. A subtle but critical point is that the ML model should only leverage input data that are available at the time when the proposed clinical decision is being made. For example, because documentation into the electronic medical record typically takes place after the encounter, using that note to detect acute stroke or myocardial infarction (where 'time is tissue') may not be of immediate clinical value.

The types of prediction tasks can be broadly categorized as learning from humans[2–5], and enabling extraction of previously unknown insights[6]. Examples of the former include facilitating large-scale screening by removing the availability and fatigue of human raters as a bottleneck, increasing the accuracy and efficiency of diagnosis. On the other end, enabling the detection of novel signals has the potential to improve diagnosis or prognosis by

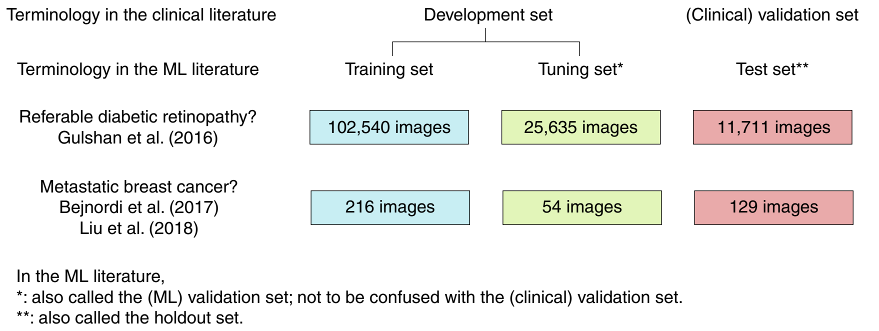

Terminology in the clinical literature
Development set
(Clinical) validation set
Terminology in the ML literature
Training set
Tuning set*
Test set**
Referable diabetic retinopathy?
Gulshan et al. (2016)
102,540 images
25,635 images
11,711 images
Metastatic breast cancer?
Bejnordi et al. (2017)
Liu et al. (2018)
216 images
54 images
129 images
In the ML literature,
*: also called the (ML) validation set; not to be confused with the (clinical) validation set.
**: also called the holdout set.

Fig. 2 | Dataset naming convention in clinical and ML studies. The figure shows the dataset naming convention using two examples, predicting referable diabetic retinopathy from retinal fundus photographs and detection of metastatic breast cancer in lymph nodes in histopathology slides.

using cheaper and scalable modalities, such as detecting cardiovascular risk via non-invasive fundus imaging[6]. Although the development and validation of both categories of predictions are similar, care must be taken to ensure that 'novel signals' found are not the result of confounding factors or random chance. To help verify the results, validation on independent datasets can be performed[7]. In addition, 'saliency' techniques[8] and qualitative assessment of the saliency 'heatmaps' can be helpful in understanding why the ML model made certain predictions. Taking this a step further, researchers have asked clinicians (who were blinded to the prediction task) to comment on the aspects of each image that was highlighted by saliency techniques, to obtain quantitative, unbiased statistics of the anatomic correlates of the predictions[6].

One key consideration in problem selection is data availability. Sufficient data are required to both develop a performant ML model and evaluate the model with high confidence that the results will generalize. For example, training a performant convolutional neural network to diagnose diabetic retinopathy from fundus images may require tens of thousands of images[2], and evaluation will require at least a thousand more. Data may be limited by the lack of digitization (such as in pathology, where most slides are not scanned), inaccessible because of patient privacy or commercial concerns, or lacking because the disease of interest is too rare. The surmountability of each of these issues varies on a case by case basis, but in general data availability presents challenges to the application of ML to healthcare. Thus, a reasonable approach may be to select problems at the intersection of useful applications and available data.

Curating datasets

The next step after problem selection is dataset construction. Because dataset naming can vary between studies, for better consistency with the clinical literature we will use 'development set' to refer to the dataset used for developing the ML model, and (clinical) 'validation set' to refer to the dataset used for final assessment of ML model performance. The development set is frequently further split into a training set, which is used to update model parameters, and a tuning set, which is used to select model hyperparameters. In the ML literature, the (clinical) validation set is frequently referred to as the test set or holdout set, while the tuning set is also called the (ML) 'validation set'. Because this can be confusing, please see Fig. 2 for an illustration. Crucially, this splitting should be 'clean' with respect to patients, for example all images from the same patient should be in the same set. When merging data from multiple sources that might involve patient-level overlap, additional effort might be needed, such as the use of image similarity to detect duplicate lesions[3].

In determining the size of the validation set, some good practices[9] from clinical trials can be helpful. Specifically, power calculations can help determine the sample size required to confidently evaluate the ML model performance or compare the model with a given baseline. In addition, all primary and secondary analyses should be pre-specified, avoiding 'post hoc' exploratory analysis. Realistically, applying ML to healthcare is frequently an exploratory process to first generate and then test hypotheses. To remain consistent with best practices for clinical trials, ML practitioners should only perform exploratory analyses on the development set, and validate the hypotheses on the validation set.

Often, the prediction task will exhibit skew in the different categories of interest, termed class imbalance. For example, if the prevalence of a disease in the population is 0.1%, on average only one example will be present for every 1,000 data points collected. The severe deficit in the number of minority class examples can limit model development and hinder accurate evaluation by enlarging the confidence intervals. Ensuring that the minority class is well-represented is also important because of the diversity of data that will be seen in real-world use, for example based on patient demographics and disease subtype. Although there are techniques to minimize the effects of class imbalance for training ML models[10], a better solution may be to augment the dataset with more examples of the minority class. This augmentation may require additional steps to ensure proper model calibration or adjustments in the evaluation metric (see section 'Evaluating model performance').

In addition, unlike ideal research settings where data can be carefully curated, real-world scenarios may require quality control to detect low-quality or irrelevant data. For example, images may have the critical diagnostic feature occluded, be out of focus or even be of the wrong anatomy (for example, images of the exterior of the eye may be provided instead of images of the retina). More subtly, the notion of 'good quality' may be disease-specific, for example the same fundus image may be of sufficient quality for assessing glaucomatous nerve head features but not for diabetic retinopathy. Depending on the specific quality issue, solutions may involve training an ML model that is robust to the issue or using a separate 'image quality' ML model[2]. For eventual real-world use, these image quality assessments would ideally be done in real time to enable on-the-spot retakes.

Another quality assessment involves the reference truth that is the target prediction of the ML model. This reference truth may be either based on the same data that the ML model sees, or an orthogonal procedure (such as a biopsy to verify radiologic findings). However, determination of the reference truth often involves subjective judgement, introducing systematic errors, random errors or both. To reduce such errors, adjudication by a panel of experts may be helpful. Higher quality labels are useful for both training more accurate models and enabling more precise estimation of the model's performance. For example, Krause et al.[4] showed that label quality has a big impact on the reliability of the evaluation metrics. The same ML model had a 30% relative reduction in errors after switching from labels established by a majority vote of three retinal specialists to labels established by adjudication from the same specialists. Fundamentally, if the reference truth is inaccurate, then the evaluation metrics will be inaccurate. Unfortunately, adjudication by a panel of experts can be slow and expensive. As such, one strategy may involve adjudication of only a subset of the data: the validation dataset for final evaluation, and the tuning dataset for hyperparameter optimization during the model development process[4].

Developing models

The process of model development for eventual clinical implementation has several main considerations that influence model architecture design: data modality and volume, model interpretability, model inference time, and balancing model overfitting and underfitting. A wide range of data modalities exist in healthcare, such as 2D images, 3D volumes, waveforms, laboratory measurements and text. For state-of-the-art performance, the ML model should be appropriate for the data modality, for example using convolutional neural networks for images[2,3,5], and recurrent neural networks for sequences of waveforms, text, measurements or text[11]. With each model type, the 'complexity' of the model (for example, as measured by the number of parameters) should also be appropriate given the dataset size. For example, a 100-layer network may overfit a classification model trained using a dataset of only 100 images.

A trend in ML is towards 'end-to-end' learning. For example, in object detection, integrating multiple models that each perform a specific 'task' into a single end-to-end model improves final prediction performance[12]. The end-to-end approach works best when large datasets are available and if the final performance is the primary metric of interest. However, in healthcare, datasets of sufficient size may be rare or absent, thus hindering direct training of end-to-end models. In addition, decomposing the model into multiple stages can have important benefits. First, the intermediate output may be useful, for example a segmentation model for tumours on pathology slides can aid review by highlighting suspicious regions even though the evaluation metric focuses on the ability to classify slides as tumour-containing or not[13,14]. Second, healthcare data can differ substantially across data sources because of factors such as institution-specific protocols and different imaging hardware. Thus, splitting a disease diagnosis step into segmentation and measurement may enable easier generalization to new imaging hardware by retraining only the segmentation model, which is more data-efficient[15]. Last, intermediate outputs may significantly enhance the interpretability of the prediction. For example, in automated cardiac measurements using ultrasound, predicting the two (x, y) coordinates that comprise the measurement 'line' may be more helpful than the final measurement[16]. On the other hand, the use of multiple stages can incur significant complexity in software implementation and model maintenance and improvements. Therefore, the decision of whether to do end-to-end learning for healthcare applications will depend on factors such as dataset size and whether the intermediate outputs are of value.

While models with a large number of parameters can have better predictive performance[17], these models also typically require more mathematical operations, and thus a longer model inference time. Particularly with the latest deep neural networks that contain tens of millions of parameters, inference can take seconds on the latest computers, precluding some models from 'real-time' (for example, >20 frames per second) usage. In addition, applications such as histopathology involve gigapixel-sized images that may require minutes to hours for inference, which may hinder intraoperative usage where time is of the essence. Fortunately, many clinical scenarios have laxer latency requirements; for example, laboratory tests may require minutes to hours. Nonetheless, the inference latency of the model should be considered during model development, and lighter-weight architectures[18] explored where appropriate.

The process of selecting a model architecture and training the model essentially involves balancing model underfitting and model overfitting, also termed the bias-variance trade-off[19]. Underfitting commonly occurs when a low-capacity model is used relative to the problem complexity and dataset size. Underfitting can be handled by selecting a more parameter-rich model or weaker regularization during training. Model overfitting is more concerning, because the evaluation overestimates the generalization performance on previously unseen data. An indication of overfitting is a surprisingly low performance on the validation set compared to that in the development set. Prevention of overfitting is typically performed by the use of a tuning set (or cross validation) within the development set for 'out of sample' estimates of performance and multiple regularization techniques (for example, parameter regularization, early stopping, pre-initialization and so on). One technique, data augmentation, is particularly useful in the data-limited healthcare regime because of its ease of application, and the ability to 'inject' prior knowledge into the training process. For example, unlike natural images, the orientation of a pathology image depends on the arbitrary orientation of the specimen during sample preparation, and thus the label (for example, tumour or non-tumour) is generally invariant to perturbations such as rotations and horizontal or vertical reflections[13]. Perturbations that may not correspond to real-world changes may still be helpful in learning, for example colour perturbations that are more extreme than those seen in the real world might still be helpful as regularization, and the 'optimal' perturbations will be based on empirical evaluation on a tuning dataset.

Finally, while developing a model, the train–tune–validation split must be carefully preserved. The validation set results should not be used for any decisions: model architecture selection, training checkpoint selection or other hyperparameter optimization. In other words, training and tuning should only be done within the development set. Because recent models tend to be high-capacity deep neural networks (that can memorize entire training datasets with random labels[20]), and tuning can have a large impact on the final performance[21], any violation of development-validation hygiene can result in ungeneralizable performance. Crucially, evaluating the performance of multiple model configurations on the validation set can result in unintended tuning using the validation set[22].

Evaluating model performance

To evaluate ML models for healthcare, the evaluation metrics have to be consistent with the ones in the relevant community. The two main categories of evaluation metrics measure discrimination and calibration[23]. Discrimination metrics measure the ability to correctly rank or distinguish two classes. The most common threshold-free discriminative metric is the area under the receiver operating characteristic curve (also called AUROC, AUC or c-statistic). Threshold-dependent metrics include sensitivity (recall), specificity and precision (positive predictive value). Thresholds tend to play a much larger role in healthcare relative to foundational ML papers because clinical applications commonly involve binary decisions, such as applying versus withholding treatment. Threshold selection depends on the clinical use case (for example, high sensitivity for screening and high specificity for diagnosis) and resource constraints (for example, only a certain percentage of patients can be screened based on time, manpower or monetary limitations).

A second class of metrics measure calibration, which evaluate how well the predicted probabilities match the actual probabilities. Some ML models do not output a probability by default and may require post-training calibration, such as Platt scaling[24]. Although under-reported, calibration metrics (for example, the Hosmer–Lemeshow statistic) are crucial for real-world use because these probabilities are used for expected cost–benefit analysis. For example, a surgery would not be indicated if the surgical risk of death is higher than the predicted probability of death without treatment. Regardless, the ML practitioner should report widely used metrics for the specific field to facilitate comparisons across studies. If no standard metrics exist, then care should be taken to report clinically relevant metrics based on the expected use case. A performant ML model should demonstrate both good discriminative performance and, where applicable, also generate well-calibrated probabilities. Furthermore, ML validation should be done using large, heterogeneous datasets to ensure generalization to diverse patient populations.

In contrast to typical ML studies, one potentially unfamiliar aspect of evaluation in healthcare involves the notion of subgroup analysis and 'population adjustment'. Better understanding of model performance in subgroups can be relevant for determining the clinical use case. In breast cancer histopathology, for example, ductal and lobular subtypes exhibit different morphologies, so additional analyses on the subtypes are needed to probe potential model weaknesses[5]. Subgroup analysis can also be based on non-patient factors, such as imaging hardware model, or the site where the data was collected. In addition, evaluation can also be affected by factors such as inclusion or exclusion of specific subgroups for the analysis. In these situations, sensitivity analysis might be prudent to ensure that these choices did not meaningfully affect the evaluation[25]. Last, for reasons mentioned above on augmenting to reduce class imbalance, the validation set collected may have a different distribution of disease subtypes relative to real-world populations. In this case, the evaluation should be adjusted according to realistic prevalence distributions[26]. This can facilitate the comparison of evaluation across literatures because the metrics won't be biased by the difference in prevalence between studies.

As in ML studies that compare the proposed method with a baseline, applications for healthcare may require comparisons with a 'human baseline' for context. For example, ML for diagnostic tasks may benefit from comparing model accuracy with that of human graders. In these situations, care should be taken to ensure a fair comparison: the experience level of the humans comprising the baseline should be representative of those in the real world, and the baseline comparators should be given a reasonable amount of time relative to real-world constraints, and they should be provided additional data such as patient history and results of other tests where relevant[27]. For ML applications that predict a previously unknown association, comparison with a baseline model (for example, logistic regression) based on variables that are readily available in the clinic (for example, demographics) may be useful to evaluate the added value of the proposed novel association.

Evaluating potential clinical impact

Developing an ML model and validating its accuracy is a necessary part of the translational process, but a performant model alone is insufficient to create clinical impact. There are many more challenges to using ML models in clinical practice because, fundamentally, no ML model will be 100% accurate in real-world scenarios. Therefore, the system has to be designed to be useful even in cases of failure such as false positives. Some issues can be handled through clinical and preclinical studies, such as selecting the most appropriate usage mode, user interface, under-reliance and over-reliance. In addition, legal and regulatory issues and implementation challenges will need to be tackled. In the process of investigating these issues, the model will need to be implemented and used in both retrospective and prospective studies, and the clinical impact measured.

First, given input data, an ML model typically provides predictions that can be used in various ways. For example, a model that assigns the diabetic retinopathy grade to a fundus image can be used 'pre-diagnosis' to pre-screen images and draw attention to images of high or uncertain risk, 'peri-diagnosis' to highlight lesions on the image during the regular image grading process[14,28], or 'post-diagnosis' to resurface images that may have been graded wrongly due to inexperience or fatigue. Each of these usage modalities may require a different user interface to present predictions in a manner that is intuitive, non-obtrusive and harmonizes with the workflow. User interface design is crucial and can mean the difference between a useful tool and a frustrating one. In detecting small metastases on a gigapixel pathology image — for example, presenting the raw predictions as a heatmap slows down readers — whereas careful filtering of the predictions to highlight only the most salient locations doubled their review speed[14].

Another key element in effectively using ML models in practice is user trust, which can result in under-reliance or over-reliance. An example of under-reliance is simply ignoring the predictions presented by the ML model, while an extreme example of over-reliance is signing off on all of the ML predictions. The degree of reliance can be measured by the user-model agreement rate, comparing that with the ML prediction accuracy[28]. The degree to which a user relies on the ML model might also be affected by factors such as whether the study affects real patient care. Thus, it is important to prospectively study the impact of using the ML model as part of an actual clinical workflow, with appropriate safeguards for patient safety. In general, user trust can be improved with clear instructions, well-designed user interface, experience in using the tool and validation studies.

Finally, the process of implementing these technologies can be complicated by factors ranging from process issues such as expensive imaging hardware or absence of software infrastructure support, to access issues such as lack of reliable internet or presence of firewalls. One often-raised issue is 'alarm fatigue', which describes the proliferation of 'alerts' in clinical workflow[29]. Because of false positives, these well-intentioned alerts may be either ignored or turned off. As such, the implementers of ML models need to consider how the new models will integrate into the workflow without undue increases of the alert burden. In addition, privacy, ethical, legal and regulatory frameworks may be locale-specific and will need to be worked through for final approval and usage in real clinical workflows. For example, a system for automated diagnosis (IDx) conducted a pre-registered, prospective 'pivotal' clinical trial that prompted FDA approval[30]. Just like drugs, however, monitoring should be done even post-approval and implementation to ensure that expected benefits materialize in different populations, and to evaluate any new harms that may arise in the course of increased and broader adoption.

Though the path towards a positive impact on patient care is a long one, we hope that these principles will streamline the process by working on important problems with the available data, develop and evaluate ML models appropriately to avoid ungeneralizable results, and pay attention to the non-ML factors (such as partnerships and user interface design) that are necessary to move towards clinical impact.

Po-Hsuan Cameron Chen[1,2]*, Yun Liu[1,2] and Lily Peng[1]
[1]*Google AI Healthcare, Mountain View, CA, USA.*
[2]*These authors contributed equally: Po-Hsuan Cameron Chen, Yun Liu.*
**e-mail: cameronchen@google.com*

Published online: 18 April 2019
https://doi.org/10.1038/s41563-019-0345-0

References

1. LeCun, Y., Bengio, Y. & Hinton, G. *Nature* **521**, 436–444 (2015).
2. Gulshan, V. et al. *JAMA* **316**, 2402–2410 (2016).
3. Esteva, A. et al. *Nature* **542**, 115–118 (2017).
4. Krause, J. et al. *Ophthalmology* **125**, 1264–1272 (2018).
5. Ehteshami Bejnordi, B. et al. *JAMA* **318**, 2199–2210 (2017).
6. Poplin, R. et al. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
7. Ting, D. S. W. & Wong, T. Y. *Nat. Biomed. Eng.* **2**, 140–141 (2018).
8. Xu, K. et al. Preprint at https://arxiv.org/abs/1502.03044 (2015).
9. Moher, D. et al. *BMJ* **340**, c869 (2010).
10. Japkowicz, N. & Stephen, S. *Intell. Data Anal.* **6**, 429–449 (2002).
11. Rajkomar, A. et al. *npj Digit. Med.* **1**, 18 (2018).
12. Ren, S., He, K., Girshick, R. & Sun, J. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2017).
13. Liu, Y. et al. *Arch. Pathol. Lab. Med.* https://doi.org/10.5858/arpa.2018-0147-OA (2018).
14. Steiner, D. F. et al. *Am. J. Surg. Pathol.* **42**, 1636–1646 (2018).
15. De Fauw, J. et al. *Nat. Med.* **24**, 1342–1350 (2018).
16. Sofka, M., Milletari, F., Jia, J. & Rothberg, A. in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (eds Cardoso, J. et al.) 258–266 (Springer, 2017).
17. Zoph, B., Vasudevan, V., Shlens, J. & Le, Q. V. Preprint at https://arxiv.org/abs/1707.07012 (2017).
18. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. in *IEEE Conference on Computer Vision and Pattern Recognition* 4510–4520 (IEEE, 2018).
19. Bishop, C. *Pattern Recognition and Machine Learning* (Springer, 2006).
20. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Preprint at https://arxiv.org/abs/1611.03530 (2016).
21. Bergstra, J. & Bengio, Y. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
22. *ILSVRC* http://www.image-net.org/challenges/LSVRC/announcement-June-2-2015 (2 June 2015).
23. Alba, A. C. et al. *JAMA* **318**, 1377–1384 (2017).
24. Niculescu-Mizil, A. & Caruana, R. in *Proc. 22nd International Conference on Machine Learning* 625–632 (ACM, 2005).
25. Thabane, L. et al. *BMC Med. Res. Methodol.* **13**, 92 (2013).
26. Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G. & Thomas, R. *Indian J. Ophthalmol.* **56**, 45–50 (2008).
27. van Smeden, M., Van Calster, B. & Groenwold, R. H. H. *JAMA* **319**, 1725–1726 (2018).
28. Sayres, R. et al. *Ophthalmology* **126**, 552–564 (2018).
29. Graham, K. C. & Cvach, M. *Am. J. Crit. Care* **19**, 28–34 (2010).
30. Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. *npj Digit. Med.* **1**, 39 (2018).
31. Shlens, J. *Google AI Blog* https://ai.googleblog.com/2016/03/train-your-own-image-classifier-with.html (2016).

Leveraging machine vision in cell-based diagnostics to do more with less

Highly quantitative, robust, single-cell analyses can help to unravel disease heterogeneity and lead to clinical insights, particularly for complex and chronic diseases. Advances in computer vision and machine learning can empower label-free cell-based diagnostics to capture subtle disease states.

Minh Doan and Anne E. Carpenter

Current diagnostic and monitoring assays are typically performed with reagents that label specific cellular and molecular hallmarks of illness (Fig. 1a). Each of these so-called biomarkers yields a single data point: the amount of the cellular constituent that has been targeted, and they often require decades of careful study to identify and validate. Furthermore, detecting biomarkers in the clinic typically requires specific reagents, special instrumentation and/or complex laboratory manipulations, which may adversely disturb the true states of the biological targets.

At the same time, the medical community recognizes that disease heterogeneity is a major challenge obscuring accurate diagnosis and effective treatment. Precision medicine, where relatively specific treatments are tailored to each patient based on their characteristics, requires new diagnostics that can classify patients by disease subtype and by response to a given treatment regime. This is particularly relevant for complex illnesses such as autoimmune diseases and cancers, and for chronic diseases that progress over time, such as diabetes and obesity. These illnesses show substantial patient-to-patient variability in terms of symptoms, genetic underpinnings and progression. Thus, a major challenge facing medicine is to develop diagnostics that reveal this heterogeneity: a given therapeutic may only be effective when it 'matches' the right patient, or even the right subclone of cells within a given patient, and response must be monitored over time.

Fortunately, cell-based diagnostics are advancing in multiple respects. Historically, cytology was limited mainly to manual microscopic examination of a biopsy specimen prepared on a slide. However, a broader range of patient phenotypes can be detected now, including multiplexed histopathology, flow cytometry, biochemistry, genetics, proteomics, immunophenotyping and more. The prospect of obtaining a more comprehensive map of disease is on the horizon.

Here we focus on several particularly promising strategies for cell-based diagnostics that identify biomarkers based on particular elements of cell morphology rather than simply the amount of a particular target molecule of the cell. They offer single-cell resolution and fine-grained classification of samples and often can be performed label-free, preserving samples' integrity and reducing the cost of reagents and instrumentation. Although not yet in widespread clinical use, we describe the advances in devices for capturing single-cell images and in machine learning that are likely to power a new generation of cell-based diagnostics, including some that are label-free.

Practical single-cell imaging platforms

Among a wide variety of single-cell analysis assays that could be employed for cell-based diagnostics, three platforms are most commonly used to detect subtle differences between individual cells using images.

Microscopy is the most common and convenient platform for imaging cells; digital cameras can capture high-resolution images of cells in multiple colorimetric and fluorescent channels, together with transmitted and phase contrast illumination. Over the centuries, many improvements in automation, illumination, photonics, optics, cameras and labelling techniques have been applied to enable image acquisition with increasing speed, resolution and specificity. Microscopy is already in widespread use for clinical cytology, where it allows microscopic examination of biopsy specimens and the inspection of the characteristic cell or tissue