A deep learning system can classify primary and metastatic cancers using passenger mutation patterns, but how does it really compare to current pathological diagnosis in a well-designed diagnostic accuracy study?

Wendy A Cooper^{1,3,4}, Laveniya Satgunaseelan^{2,3}, Ruta Gupta^{1,3}

¹*Department of Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital and NSW Health Pathology, Sydney, NSW, Australia*

²*Department of Neuropathology, Royal Prince Alfred Hospital*

³*Sydney Medical School, University of Sydney, Sydney, NSW, Australia*

⁴*School of Medicine, Western Sydney University, Campbelltown, NSW, Australia*

Competing interest statement: All authors declare no conflict of interest, nor any competing interests.

Author contribution statement: All authors contributed extensively to the work presented in this paper.

ABSTRACT:

In a recent study published in *Nature Communications* by Jiao W et al, a deep learning classifier was trained to predict cancer type based on somatic passenger mutations identified using whole genome sequencing (WGS) as part of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium. The data show patterns of somatic passenger mutations differ between tumours with different cell of origin. Overall, the system had an accuracy of 91% in a cross-validation setting using the training set, and 88% and 83% using external validation sets of primary and metastatic tumours respectively. Surprisingly, this is claimed to be twice as accurate as trained pathologists, based on a 27 year old reference from 1993 prior to availability and routine utilisation of immunohistochemistry (IHC) in diagnostic pathology and is not a reflection of current diagnostic standards. We discuss the vital role of pathology in patient care and the importance of using international standards if deep learning methods are to be used in the clinical setting.

In the study by Jiao W et al (1), a deep learning classifier was trained to predict cancer type based on somatic passenger mutations identified using WGS as part of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium. The data show patterns of somatic passenger mutations differ between tumours with different cell of origin. Overall, the system had an accuracy of 91% in a cross-validation setting using the training set, and 88% and 83% using external validation sets of primary and metastatic tumours respectively. Surprisingly, this is claimed to be twice as accurate as trained pathologists assessing a metastatic tumour without given knowledge of the primary. This assertion is based on a 27 year old reference from 1993 (2) prior to availability and routine utilisation of immunohistochemistry (IHC) in diagnostic pathology and is not a reflection of current diagnostic standards.

A more recent diagnostic accuracy study from 2013 comparing standard pathological evaluation with gene expression profiling (GEP) in metastatic tumours with known primaries, found pathologists had 83% accuracy (3). This is the same as that reported by Jiao et al using a deep learning approach. Interestingly, GEP had a slightly superior accuracy of 89%. In cases where pathologists only required a single round of IHC, the accuracy of pathologists and GEP was >90%. Pathologist accuracy varied by tumour site and was lower in poorly differentiated/undifferentiated tumours. If comparisons of diagnostic accuracy are to be based on literature review, then it is critical that studies reflecting current clinical practices are included.

Using the QUADAS-2 tool for quality assessment of diagnostic accuracy studies, which includes parameter such as patient selection, index test, reference standard and flow and timing, the study by Jiao et al is judged as “at risk of bias” and having “concerns regarding applicability” (4). There are several reasons for this bias. For example, studies of diagnostic accuracy should follow STARD guidelines to ensure essential factors such as eligibility criteria, reasons for exclusion of cases, handling of indeterminate results are reported and that assessment of bias can be under taken and this is usually a requirement for publication (5).

Cases included in the study should be representative of the study question. Some tumours are “so poorly differentiated” that they are difficult to classify based on morphology and IHC, and it would be interesting to assess the genotype of these tumours to determine if NGS and deep machine learning methods could help determine their cell of origin. The study by Jiao et al include 62 CUPs but does not clarify how many are poorly differentiated tumours that are difficult to classify. Interestingly, similar to current diagnostic practices the deep learning approach provided up to 3 differential diagnoses with probability estimates. It is uncertain how many of these had pathologic evaluation for clinical management that may have similarly provided differential diagnoses. It is also difficult to assess the accuracy of the classifier without being aware of the clinical and radiologic details. For example, did all of the patients in whom hepatocellular carcinoma was suggested as a primary have

liver lesions? In addition to common adenocarcinomas (e.g. breast, colorectal, prostate, lung, pancreatic), the study by Jiao et al includes several brain tumours (pilocytic astrocytoma and “diffuse glioma”) and haematological malignancies (chronic lymphocytic leukaemia and “myeloproliferative neoplasm”). These would essentially never be of any diagnostic confusion with carcinomas based on their completely distinct morphological appearances by light microscopy (Figure 1). The majority of poorly differentiated malignant neoplasms can also be accurately classified using IHC (Figure 2).

Sensitivity analyses form a critical part of diagnostic accuracy and should also include cases that are non-evaluable such as cases with insufficient DNA quantity or quality should be included as non-concordant cases. This is particularly important in the context of carcinoma of unknown primary, as a very limited biopsy is generally available for clinical diagnostic assessment. A widely reproducible reference standard should be utilised but it is clear from the diagnoses provided in this study that a consistent pathological classification system such as the WHO classification of tumours was not utilised in all cases.

Jiao et al state that histological and immunohistochemical analysis of tumours can be “complex and time consuming”. The average turn-around times for pathological light microscopic assessment of tumours is few days from the time of biopsy and can be expedited to 24 hours if clinically urgent. IHC is a low cost and rapid molecular technique taking <24h enabling clinical treatment decisions to be made in a timely manner. While the costs and time required for whole genome sequencing are significantly reducing, these are yet to match the cost effectivity and turnaround time of histochemical and IHC widely available to all patients.

Adequately designed studies of diagnostic accuracy that account for contemporary clinical and diagnostic practices, cost effectivity, clinically relevant turnaround times and using internationally standardised analytic pipelines are needed if WGS and deep learning methods are to be used as a “diagnostic tool for tumour-type identification”. The study needs to utilise appropriate cohorts and adhere to STARD reporting guidelines before claims about diagnostic accuracy or “clinical applicability” can be made. Any diagnostic technique that outperforms current approaches and that has demonstrated clinical utility are welcome to improve patient care but misleading studies that do not provide evidence to support their claims should be discouraged. The study by Jiao et al is hypothesis generating, however, the authors have claimed high performance and clinical significance beyond what has been proven.

References:

1. Jiao W, Atwal G, Polak P, Karlic R, Cuppen E, Danyi A, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature communications* 2020;11:728.
2. Sheahan K, O'Keane JC, Abramowitz A, Carlson JA, Burke B, Gottlieb LS, et al. Metastatic adenocarcinoma of an unknown primary site. A comparison of the relative contributions of

morphology, minimal essential clinical data and CEA immunostaining status. American journal of clinical pathology 1993;99:729-35.

3. Handorf CR, Kulkarni A, Grenert JP, Weiss LM, Rogers WM, Kim OS, et al. A multicenter study directly comparing the diagnostic accuracy of gene expression profiling and immunohistochemistry for primary site identification in metastatic tumors. The American journal of surgical pathology 2013;37:1067-75.

4. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine 2011;155:529-36.

5. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ (Clinical research ed) 2015;351:h5527.

Figure legend:

1. Histopathological examples of well differentiated tumours: (A) lung adenocarcinoma (haematoxylin and eosin (H & E) staining 40X); (B) pilocytic astrocytoma (H & E, 40X); (C) small lymphocytic lymphoma (H&E, 40X).
2. Histopathological examples of poorly differentiated tumours with corresponding panel of widely used immunohistochemical stains: (A) poorly differentiated lung carcinoma (H&E, 40X) with (B) positive cytokeratin immunohistochemistry; (C) diffuse large B-cell lymphoma (H&E, 40X) with (D) positive CD20 immunohistochemistry; (E) melanoma with (F) positive S100 immunohistochemistry.