

Article

Automated PD-L1 Scoring for Non-Small Cell Lung Carcinoma Using Open-Source Software

Naso JR, Povshedna T, Wang G, et al. Pathol Oncol Res. 2021 Mar 26;27:609717.From Canada

はじめに：非小細胞肺癌に対するコンパニオン診断としてPD-L1の発現を調べることはほぼ日常化している。発現率によって治療方針が変わるため正確な測定を必要とするが、診断者間の差が生じることが問題とされる。筆者らはデジタル画像を用いた自動計測がPD-L1発現状態をより正確に測定できるのではないかと考え、無料の画像解析ソフトのQuPathを用いてスコアを算出し3人の病理医の平均スコアとの一致率を評価した論文である。

TABLE 1 | Case demographics.

Category	Training set		Testing set	
	n	%	n	%
Sex				
Female	5	50%	35	51%
Male	5	50%	34	49%
Diagnosis				
Adenocarcinoma, non-mucinous	8	80%	58	84%
Squamous cell carcinoma	1	10%	11	16%
Non-small cell carcinoma NOS	1	10%	0	0%
Site				
Lung	10	100%	66	94%
Metastasis to bone	0	0%	1	1%
Metastasis to brain	0	0%	2	3%
Procedure				
Lobectomy	6	60%	45	65%
Wedge resection	4	40%	21	30%
Large resection of metastatic tumor	0	0%	3	4%
Image patch area [mm ² , median (range)]	0.94 [0.16–7.94]	NA	0.44 [0.05–4.01]	NA

PD-L1免疫組織化学が行われた非小細胞肺癌各症例から3つの代表的なImage patchを病理医が選
び、デジタル画像化しQuPathにて解析を行った。まず10症例、30のImage patchを用いてトレーニ
ングを行った。腫瘍細胞とそれ以外の細胞をわけるため、67項目にわたる形態的な特徴が1個1個
の細胞で計算され、PD-L1発現を計測するためのclassifierが作製された。それを用いて69症例、

207個のImage patchの解析が行われた。比較対象として3人の病理医が発現状態を計測し、その平均値を本研究のgold standardとした。発現率は0-100%の間で計測されるが、最初は発現率そのものの数値を算出し、次に3つのカテゴリー（<1%, 1-49%, and ≥50%）にわけて検討された。

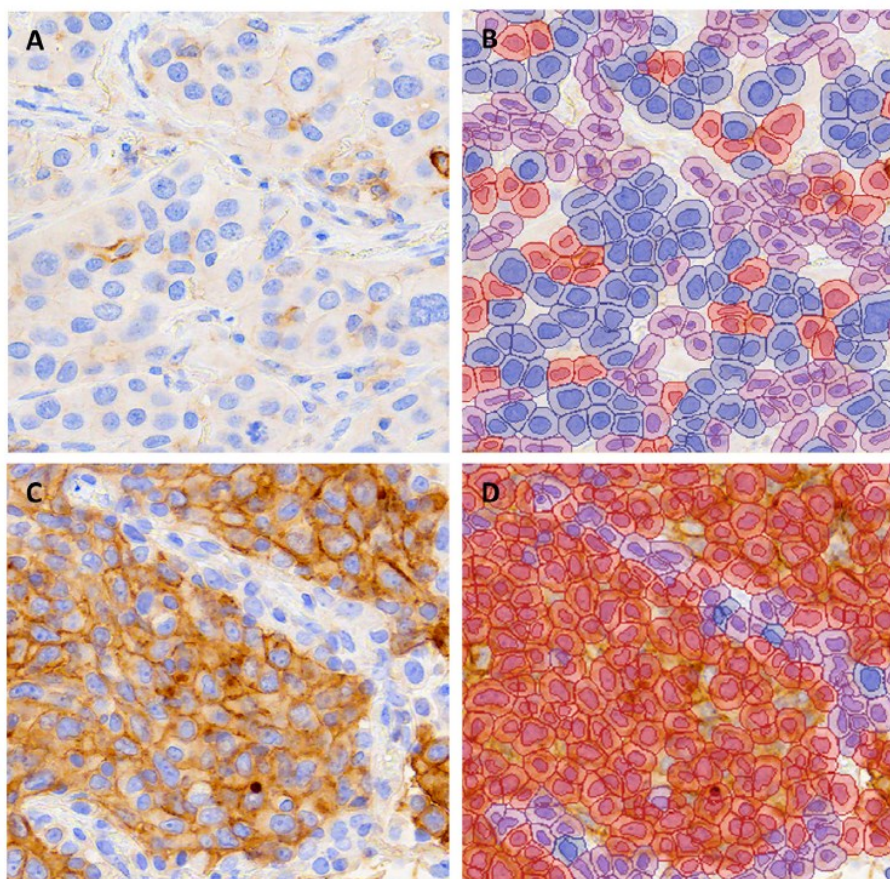


FIGURE 1 | Representative images of PD-L1 stained non-small cell lung carcinomas from (A,B) a case with TPS 1–49% and (C,D) a case with TPS ≥50% in all scoring methods (~X200). The unannotated image is shown in (A,C) and QuPath annotations are shown in (B,D). Blue indicates PD-L1 negative tumor cells, red indicates PD-L1 positive tumor cells and purple indicates background stroma and immune cells.

TABLE 2 | Agreement of automated scores and individual pathologist's manual scores with the average manual score for single image patches.

Threshold	Scoring method	Concordant with average manual score (n [%])	Cohen's kappa for agreement with average manual score (95% CI)	Sensitivity ^a (%)	Specificity ^a (%)
1	QuPath	186/207 (90%)	0.80 (0.71–0.88)	95	84
	Pathologist #1	188/207 (91%)	0.82 (0.74–0.89)	83	100
	Pathologist #2	191/207 (92%)	0.85 (0.77–0.92)	87	98
	Pathologist #3	197/207 (95%)	0.90 (0.84–0.96)	98	92
50	QuPath	191/207 (92%)	0.78 (0.69–0.92)	71	100
	Pathologist #1	200/207 (97%)	0.91 (0.85–0.98)	95	97
	Pathologist #2	196/207 (95%)	0.87 (0.80–0.95)	98	93
	Pathologist #3	201/207 (97%)	0.93 (0.87–0.98)	93	99

^aThe average manual score was used as the 'gold standard' for sensitivity and specificity calculations.

まず計測された数値で検討すると、自動計測はgold standardと強い相関（coefficient 0.925）を示した。次にカテゴリー別にわけて検討、1%以上陽性をthresholdとすると一致率は186/207(90%), 50%以上陽性とする一致率は191/207(92%)であり、kappa係数は1%以上で $\kappa=0.8$, 50%以上で0.78と強い相関をしめし、個々の病理医とgold standardのkappa係数と差はなかった。

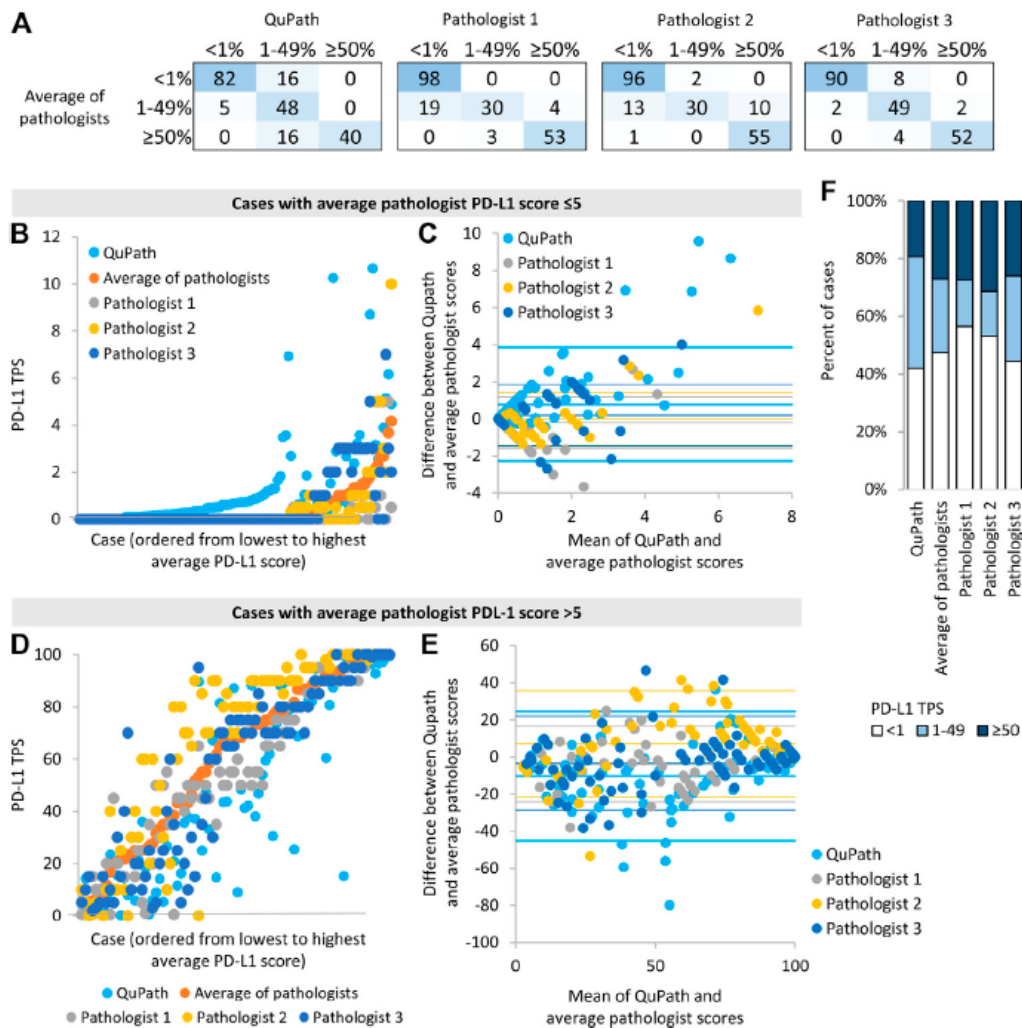


FIGURE 2 | Comparison of automated (“QuPath”) and manual (“pathologist”) PD-L1 scores **(A)** The number of image patches in each PD-L1 TPS category when using scores from automated or single pathologist scoring compared to average pathologist scores **(B,D)** Scores for each image patch ordered by the average pathologist score **(C,E)** Bland-Altman plots of the difference in scores vs. the mean of scores for different scoring methods compared to average pathologist scores. For each data series, the horizontal lines represent two standard deviation above the mean difference, the mean difference, and two standard deviations below the mean difference. Data for image patches and data for image patches with an average score ≤5% are shown in **(B,C)**, and those with an average pathologist score >5% are shown in **(D,E)**. **(D)** The proportion of image patches in each PD-L1 score category according to each scoring method.

次に自動計測がgold standardと比べて、スコアをどのくらい過剰評価または過小評価しているのかを検討した。1%をthresholdとして両者を見てみると、16のpatchが自動計測で1-49%、gold standardで<1%であったのに対し、5つのpatchが自動計測で<1%, gold standardで1-49%であった。（Figure 2A）数値で見ると

gold standardで5%以下だった症例は自動計測では平均0.8%gold standardより高値を示した。(Figure 2B,C) そして95%の感度、84%の特異度であった。感度は個々の病理医よりも高く、特異度は低かった (Table 2参照)。対照的に、50%をthresholdとして両者を見てみると、16のpatchが自動計測で1-49%、gold standardでは $\geq 50\%$ であった。(Figure 2A) 数値で見るとgold standardで5%を超えた症例は自動計測では平均10%gold standardより低値を示した。Figure 2D,E) そして71%の感度、100%の特異度であった。特異度はいいが、感度は個々の病理医よりも低かった。(Table2参照) 自動計測が高スコアの症例を過小評価または低スコアの症例を過剰評価するのを反映するように、自動計測では1-49%のpatch数が他と比べて多かった。(Figure 2F) これは自動計測で1-49%とされたpatchは不一致になりやすいことを示している。

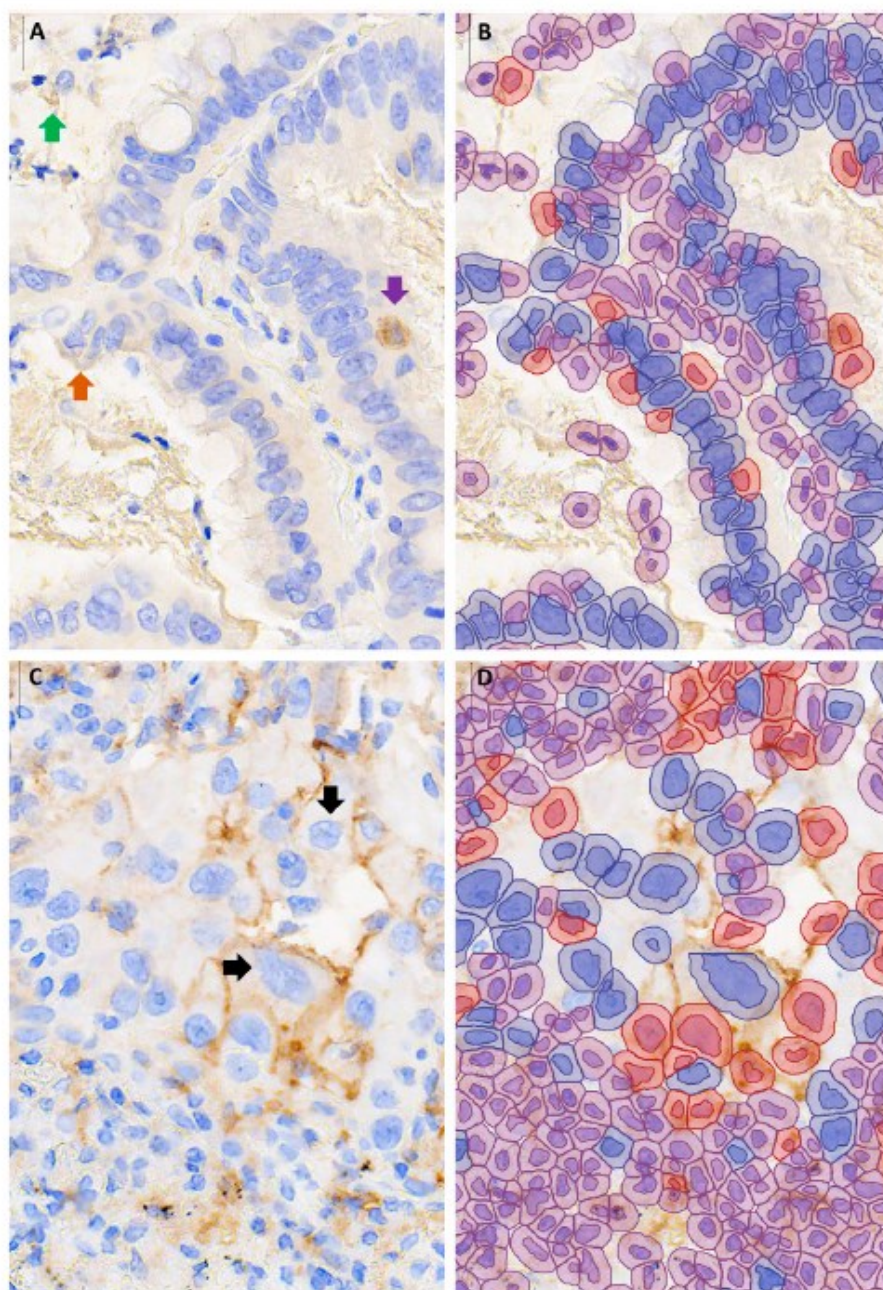


FIGURE 3 | Examples of cases with discordant automated and manual PD-L1 scores. QuPath annotations (**B,D**) are shown to the right of the corresponding unannotated PD-L1 stained area, 400–600X magnification (**A,C**). (**A,B**) Over-estimation of PD-L1 staining using QuPath was contributed to by false positive annotation of debris (green arrow), inflammatory cells (purple arrow) and edge artifact (orange arrow). (**C,D**) Under-estimation of PD-L1 staining using QuPath was contributed to by cell area being underestimated, such that membranous staining fell outside of the annotated area of the cell ↓.

不一致であった37patchについてみると、11patchは少なくとも1人の病理医のスコアと一致していた。

13patchは自動計測でgold standardより高く計測されたが、そのほとんどは2%未満であった。上記

Figure3A,Bは自動計測で高く計測された症例である。Debris(緑矢印)、炎症細胞（紫矢印）、edge-artifact

（オレンジ矢印）の部分が陽性と計測された。一方で13patchは自動計測で低く計測された。そのうち

9patchは自動計測で1-49%のカテゴリーに区分された。低く計測された要因としては非常に弱い染色性、膜染色がannotationの外側にあるため陰性とされた（Figure 3C,D）などである。

TABLE 3 | Agreement of automated scores and individual pathologist's manual scores with the average manual score for each patient case.

Threshold (%)	Scoring method	Concordant with average manual score (n [%])	Cohen's kappa for agreement with average manual score (95% CI)	Sensitivity ^a (%)	Specificity ^a (%)
1	QuPath	64/69 (93%)	0.77 (0.62–0.93)	98	85
	Pathologist #1	63/69 (91%)	0.84 (0.71–0.97)	88	96
	Pathologist #2	67/69 (97%)	0.94 (0.86–1)	95	100
	Pathologist #3	65/69 (94%)	0.87 (0.76–0.99)	98	88
50	QuPath	63/69 (91%)	0.76 (0.58–0.94)	68	100
	Pathologist #1	68/69 (99%)	0.96 (0.89–1)	95	100
	Pathologist #2	67/69 (97%)	0.93 (0.83–1)	100	96
	Pathologist #3	67/69 (97%)	0.93 (0.83–1)	95	98

^aThe average manual score was used as the 'gold standard' for sensitivity and specificity calculations.

より実臨床的にするために、3patch/1 caseの重み付き平均値を計算し、69症例において自動計測とgold standardとの比較検討を行った。結果としては先に検討したsingle patchの結果とほぼ同様であり、より大きな組織に対しても適用されると考えられた。

Take home message

- ・ QuPathを用いたPD-L1スコアリングは実臨床にも使用可能と考えられた。
- ・ より正確性を追求することにより、病理医の仕事軽減になる可能性が考えられた。

（文責：柳川直樹）