Improving the Reliability of Medical Diagnostic Models through Rule-Based Decision Deferral

Jacqueline Isabel Bereska [1, 2, 3], Henk Marquering [1, 2, 3], Marc Besselink [2, 4], Jaap Stoker [1, 2] Inez Verpalen [1, 2]

[1]Amsterdam UMC, location University of Amsterdam, Department of Radiology and Nuclear Medicine, Amsterdam, The Netherlands

[2]Cancer Center Amsterdam, Amsterdam, The Netherlands

[3] Amsterdam UMC, location University of Amsterdam, Department of Biomedical Engineering and Physics, Amsterdam, The Netherlands

[4] Amsterdam UMC, location Free University of Amsterdam, Department of Surgery, Amsterdam, The Netherlands

{j.i.bereska, h.a.marquering, m.g.besselink, j.stoker, i.m.verpalen}@amsterdamumc.nl

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is a highly lethal cancer, and accurate assessment of tumor resectability is crucial for determining appropriate treatment. AI-based models have shown promise in classifying tumor resectability, but reliability concerns have impeded clinical implementation. We propose extending the AI-based VasQNet model for classifying tumor resectability on AI-generated segmentations of computed tomography scans (CTs) to improve the models' reliability. This extension allows VasQNet to defer decisions when the AI-generated segmentations violate pre-established rules on vascular anatomy, tumor location, and tumor size.

We conducted experiments using CTs of (borderline) resectable and non-resectable PDAC patients. We evaluated the performance of the baseline VasQNet and the extended VasQNet with rule-based decision deferral (RBDD) by comparing their classifications to a ground-truth provided by a radiologist, employing agreement as a metric.

Our results demonstrate that the extended VasQNet achieved a significantly higher agreement (90%) with the radiologist's classification than the baseline VasQNet (67%). Notably, 17/31 (54%) deferred decisions would have been incorrect had they not been deferred. Our study demonstrates the effectiveness of RBDD in improving the reliability of clinical diagnostic models through the exemplification of VasQNet. In conclusion, RBDD can enhance the reliability of clinical diagnostics models, facilitating integration into clinical practice.

The documented code is available on GitHub (https://github.com/PHAIR-Consortium/Vessel-Involvement-Quantifier).

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is a highly lethal cancer with a low five-year survival rate (Chhoda et al. 2019). Surgical tumor resection combined with systemic therapy is currently the only curative treatment option. PDAC resectability, which heavily depends on the degree of vascular involvement, can be challenging to assess. Radiologists typically use contrast-enhanced computed tomography scans (CTs) to assess the degree of vascular involvement, but the high degree of interobserver variability can lead to suboptimal patient selection for surgery and clinical studies (Giannone et al. 2021). Recent efforts to address this issue have focused on AI-based models to assess vascular involvement and tumor resectability (Rigiroli et al. 2021; Bian et al. 2020; Lao et al. 2020; Chen et al. 2020). These models hold promise in providing more accurate and standardized assessments. However, their clinical implementation is hindered by the lack of reliability and interpretability of their assessments (Lee et al. 2020; Aristidou, Jena, and Topol 2022).

Most AI-based models for assessing vascular involvement and tumor resectability use radiomic features, which are not readily interpretable to clinicians, to make assessments (Rigiroli et al. 2021; Bian et al. 2020; Lao et al. 2020; Chen et al. 2020). To address this limitation, VasQNet, a model that combines established clinical guidelines and AI-based segmented anatomical structures, was proposed (Bereska et al. 2023). By mirroring the current radiological workflow, VasQNet provides easier interpretable assessments to clinicians than earlier radiomics-based models. However, indiscriminate classification approaches generating classifications for every case, regardless of their certainty, still prevail in most AI-based models, including VasQNet. This indiscriminate classification approach forces clinicians to double-check every classification thoroughly, reducing the model's utility. This is especially crucial when the model's decision is critical, including determining a patient's eligibility for surgery. Erroneous classifications in such cases could result in devastating consequences, such as excluding eligible patients from surgery or performing surgery on patients with advanced PDAC, adversely impacting treatment outcomes and quality of life. Therefore, a reliable model that generates classifications only when reasonably certain and defers classifications otherwise is urgently needed.

Several techniques, such as uncertainty quantification, have been proposed to enable decision deferral (Joshi, Parbhoo, and Doshi-Velez 2021; Liu, Gallego, and Barbieri 2022). However, these techniques rely on a probabilistic model framework, which does not apply to deterministic models such as VasQNet, as they yield non-probabilistic

classifications. To overcome this limitation, we propose an extension to VasQNet that defers decisions when the underlying AI-generated segmentations of anatomical structures violate a set of pre-established rules that are either known or likely to be true. This approach aims to mitigate erroneous classifications, enhance the model's reliability, and ultimately facilitate clinical translation.

Methods

Datasets

We included two datasets, *A* and *B*, comprising 60 late arterial phase CTs obtained from 60 patients diagnosed with resectable, borderline resectable, and non-resectable PDAC. Dataset *A* represents the subset of the PREOPANC trials conducted by the Dutch Pancreatic Cancer Group (DPCG) at the Amsterdam UMC between 2013 and 2020. Dataset *B* consists of patients from the DPCG non-resectable PDAC registration at the Amsterdam UMC between 2019 and 2021. At the time of the CT scan, a specialized abdominal radiologist assessed the vascular involvement and tumor resectability for each patient. These assessments were subsequently discussed at the multidisciplinary oncology meeting and stored in the Picture Archiving and Communication System (PACS) of the hospital. General informed consent was obtained from all patients.

Baseline VasQNet Model

The workflow of the baseline VasQNet model comprises three steps: 1) automatic segmentation of PDAC and neighboring vessels, 2) quantification of vascular involvement, and 3) classification of tumor resectability.

VasQNet was trained to segment the PDAC and surrounding vasculature, specifically the aorta, celiac trunk (CeTr), hepatic artery (HA), superior mesenteric artery (SMA), superior mesenteric vein (SMV), and portal vein (PV), using a self-learning approach. The self-learning approach consists of training a teacher segmentation model on a small amount of manually annotated data and subsequently using the resulting teacher model to provide annotations for a large dataset which is then used to train a student segmentation model.

To quantify the degree of vascular involvement of the CeTr, HA, SMA, SMV, and PV by the PDAC tumor, VasQNet evaluates all x, y, and z dimensions containing both the vessel and the tumor. For each vessel, VasQNet calculates the circumference of the vessel segment (V) and the length of its connection with the tumor segment (TV) in each x,y, and z plane. The degree of involvement for each dimension is then determined using the formula $\frac{TV}{V} \cdot 360$. The maximum degree of involvement across all dimensions is selected for each vessel, resulting in a continuous variable ranging from 0 to 360 degrees.

VasQNet determines tumor resectability by considering the degrees of vascular involvement for each vessel and referring to the DPCG resectability guidelines, which classify tumors as resectable, borderline resectable, or non-resectable based on the criteria outlined in Table 1.

Decision Deferral Model

To enhance the reliability of the baseline VasQNet model, we integrated three rules rooted in established clinical expertise pertaining to vascular anatomy, tumor location, and tumor size into the classification process.

Vascular Anatomy Given the variability in contrast administration and the utilization of arterial phase scans over venous phase scans, accurate segmentation of vessels, particularly veins, is occasionally compromised by VasQNet. To ensure that predictions are not generated based on CTs with substantial segmentation errors, we introduce anatomy-based rules that align with established anatomical verities. Specifically, VasQNet will defer the classification if the AI-generated segmentations violate the following anatomical rules: 1) the TC does not originate from the aorta, 2) the HA does not originate from the TC, 3) the SMA does not originate from the aorta, and 4) the SMV and the PV are not connected.

Tumor Location Extrapancreatic tumor growth leads to closer proximity and direct contact with blood vessels in the surrounding anatomical region. This close association facilitates the tumor's interaction with the vasculature, resulting in a higher probability of (extensive) vascular involvement with adjacent vessels. Besides increasing the likelihood of vascular involvement, extrapancreatic tumor growth also can also result in vascular infiltration or compression, causing occlusion, narrowing, or compromised blood flow, further necessitating diligent scrutiny of these scans by radiologists (Toshima et al. 2022). Hence, in resectable cases where less than 25% of the periphery of the tumor segment calculated on all x dimensions is in contact with the pancreas, the model will defer the classification.

Tumor Size Increased tumor size is associated with a higher likelihood of vascular involvement due to the tumor's greater physical mass and potential for local infiltration. As the tumor grows, it can exert mechanical pressure on the surrounding blood vessels, leading to vessel infiltration, compression, and increased chances of vascular involvement, emphasizing the significance of considering tumor size in assessing vascular involvement (Li et al. 2018). Past research has demonstrated an association between PDAC tumor diameters exceeding 20 mm and poor prognosis (Marchegiani et al. 2017). To account for these factors, the model will defer classification for resectable tumors with a diameter exceeding 20 mm on the AI-generated segmentations.

Performance Assessment

To evaluate the performance of both the baseline VasQNet and the extended model in terms of tumor resectability, we compared their classifications with the ground-truth classification obtained from the single radiologist assessment stored in the PACS system, using agreement as the metric. Agreement measures the number of cases in which the model and the single radiologist yielded the same tumor resectability classification. A chi-squared test was conducted on the contingency table of the two models' agreement to assess the

Category	Celiac trunk	Hepatic artery	Superior mesenteric artery	Superior mesenteric vein	Portal vein
Resectable	0°	0°	0°	0 – 90°	0 – 90°
Borderline Resectable	0 – 90°	0 – 90°	0 – 90°	0 – 270°	0 – 270°
Non-resectable	90 – 360°	90 – 360°	90 – 360°	270 – 360°	270 – 360°

Table 1: Guidelines of the Dutch Pancreatic Cancer Group for classifying PDAC resectability based on the involvement with five vessels (Dutch Pancreatic Cancer Group (DPCG) 2023).

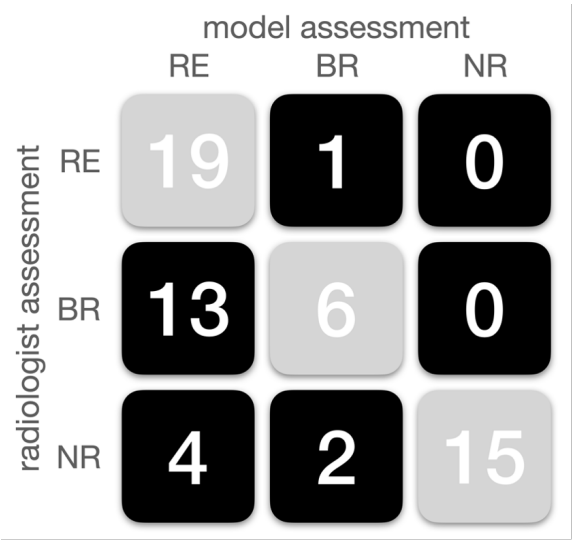

Figure 1: Agreement of baseline VasQNet with radiologist's assessment for classifying tumor resectability for resectable (RE), borderline resectable (BR), and non-resectable (NR) PDAC.

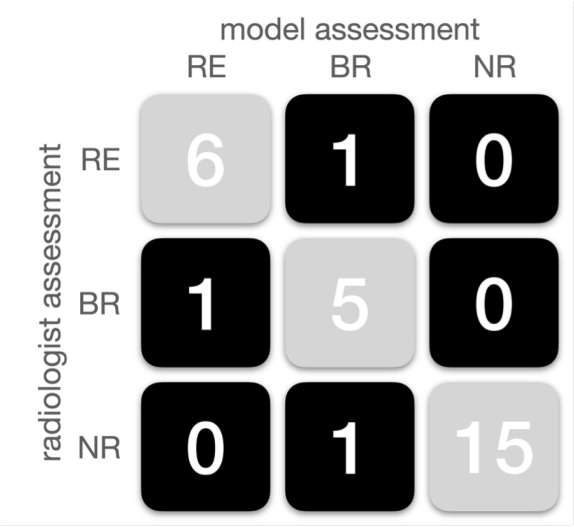

Figure 2: Agreement of VasQNet with rule-based decision deferral with radiologist's assessment for classifying tumor resectability for resectable (RE), borderline resectable (BR), and non-resectable (NR) PDAC.

statistical significance of the observed difference in agreement. A p-value less than 0.05 was deemed statistically significant.

Results

Patient Characteristics

The test set comprised 60 late arterial phase CT scans obtained from 60 patients, including 20 patients with resectable PDAC, 20 patients with borderline resectable PDAC, and 20 patients with non-resectable PDAC. Among the patients in the test set, 32 (53%) were females, and the median tumor diameter was 3.4 cm with a standard deviation of 1.7 cm.

Classifying Resectability with the Baseline VasQNet Model

The baseline VasQNet model classified tumor resectability with an agreement of 40/60 (67%). The agreement between the radiologist and VasQNet for classifying tumor resectability is illustrated in the agreement matrix shown in Figure 1.

Classifying Resectability with Rule-Based Decision Deferral

By deferring decisions when any of the three rules pertaining to vascular anatomy, tumor size, or location were violated, we classified tumor resectability with an agreement of 26/29 (90%). The agreement between the radiologist and VasQNet extended with rule-based decision deferral for classifying tumor resectability is illustrated in the agreement matrix shown in Figure 2. Overall, 31/60 (52%) of the AI-generated segmentation violated one or more of the three rules resulting in decision deferral. We found a statistically significant difference in agreement between the baseline VasQNet model and the extended model (p-value < 0.001).

In total, 17 of the 31 deferred decisions (53%) would have been incorrect had they not been deferred. Specifically, 17 decisions were deferred due to tumor size, while nine decisions were deferred based on tumor location. Moreover, three decisions were deferred as a result of both tumor size and deviating vascular anatomy, and an additional two decisions were deferred due to the combined factors of tumor size, tumor location, and deviating vascular anatomy.

Discussion

This study examined rule-based decision deferral to enhance the reliability of clinical diagnostics models, using VasQNet as an example. We observed a significant reduction in false negative and false positive classifications by comparing the baseline VasQNet model with an extended version incorporating rule-based decision deferral. This finding highlights

the potential of rule-based decision deferral as a valuable tool for improving the reliability of clinical diagnostics models.

Previous research in uncertainty quantification has primarily focused on probabilistic methods to estimate and quantify uncertainty in model decisions. These approaches typically rely on statistical techniques, Bayesian inference, or Monte Carlo simulations to provide probabilistic assessments of decision reliability (Zhang 2021; Kwon et al. 2018; Taghizadeh, Karimi, and Heitzinger 2020; Abdar et al. 2021a,b). While uncertainty quantification methods have demonstrated valuable insights, they often suffer from computational complexity, limited interpretability, and the need for large amounts of data for accurate estimation. In contrast, rule-based decision deferral offers several advantages. It provides clinicians with explainable insights that align with their existing workflow and are easily interpretable. Rule-based decision deferral can be applied to a wide range of clinical diagnostics models and is compatible with non-probabilistic models and observations as well as small data sets. Additionally, the flexibility of rule-based decision deferral enables the easy extension of rules without retraining the underlying model, facilitating adaptability and continuous improvement in clinical decision-making.

There are several limitations to consider when interpreting our research. First, it is crucial to acknowledge that predefined rules represent simplifications and may not encompass the full range of possible scenarios. For example, patients with PDAC can present with atypical anatomical characteristics or exhibit features such as large tumors with extrapancreatic growth, even in the absence of vascular involvement. Consequently, a notable limitation of the rule-based approach is its inherent tendency to also defer decisions that would have been accurate if they had not been deferred. Nevertheless, it is important to note that even in cases where the deferred decisions may have been correct, the presence of factors such as deviating anatomies introduces additional complexities or may necessitate alternative surgical approaches, thereby warranting careful consideration irrespective of resectability. Second, the rules presented in this paper are specific to PDAC and may not be directly applicable to other pathologies. Other pathologies likely require conceptualizing a new set of rules which subsequently need to be validated. In our study, the selection of pre-established rules for decision deferral was based on established anatomical considerations that were either known or presumed. However, it is essential to recognize that the chosen rules may not encompass all relevant factors that influence the assessment of resectability. Therefore, continuous refinement and updating of the rules based on emerging clinical evidence is imperative to ensure the optimal performance of the model. Lastly, due to the availability of only one assessment per patient in the PACS system, the method was compared against a single radiologist assessment. However, it is important to note that the radiologists differ between the patients, reducing the likelihood of bias from a single observer.

Lastly, the method was only compared against the single radiologist assessment in the PACS system. However, it is important to note that the assessments are provided by different radiologists for each patient. Furthermore, these assessments undergo thorough discussion and evaluation at MDO meetings, ensuring comprehensive analysis.

Several avenues for future research can further enhance the application of rule-based decision deferral in clinical diagnostics models. First, efforts can be directed towards automating the process of rule generation by leveraging machine learning techniques. This would involve applying algorithms that can automatically identify and extract relevant rules from a training dataset, thereby reducing the reliance on manual rule specification. Additionally, investigating the impact of rule-based decision deferral on clinicians' trust in diagnostics models would yield valuable insights into the acceptance and adoption of these models in clinical practice. Furthermore, conducting prospective validation studies encompassing larger and more diverse patient cohorts would provide crucial evidence regarding the generalizability and clinical effectiveness of this approach. Lastly, it is imperative to continuously refine and update the rules based on emerging clinical evidence and expert consensus, ensuring the ongoing accuracy and applicability of the decision deferral approach in real-world healthcare settings.

Conclusion

Our study demonstrates that incorporating rule-based decision deferral in clinical diagnostics models such as VasQNet improves reliability by significantly reducing false positives and negatives. By aligning with established clinical knowledge and mimicking the decision-making process of radiologists, this approach may facilitate clinician understanding and promote the integration of AI systems into clinical practice. Future research should focus on prospective validation and implementation studies to further evaluate the clinical impact and feasibility of rule-based decision deferral in real-world settings.

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; et al. 2021a. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76: 243–297.

Abdar, M.; Samami, M.; Mahmoodabad, S. D.; Doan, T.; Mazoure, B.; Hashemifesharaki, R.; Liu, L.; Khosravi, A.; Acharya, U. R.; Makarenkov, V.; et al. 2021b. Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. *Computers in biology and medicine*, 135: 104418.

Aristidou, A.; Jena, R.; and Topol, E. J. 2022. Bridging the chasm between AI and clinical implementation. *The Lancet*, 399(10325): 620.

Bereska, J. I.; Janssen, B. V.; Nio, C. Y.; Kop, M. P. M.; Kazemier, G.; Busch, O. R.; Struik, F.; Marquering, H. A.; Stoker, J.; Besselink, M. G.; and Verpalen, I. M. 2023. Automatic assessment of tumor resectability on computed tomography in patients with localized pancreatic cancer using artificial intelligence. Forthcoming.

Bian, Y.; Jiang, H.; Ma, C.; Cao, K.; Fang, X.; Li, J.; Wang, L.; Zheng, J.; and Lu, J. 2020. Performance of CT-based radiomics in diagnosis of superior mesenteric vein resection margin in patients with pancreatic head cancer. *Abdominal Radiology*, 45: 759–773.

Chen, F.; Zhou, Y.; Qi, X.; Zhang, R.; Gao, X.; Xia, W.; and Zhang, L. 2020. Radiomics-assisted presurgical prediction for surgical portal vein-superior mesenteric vein invasion in pancreatic ductal adenocarcinoma. *Frontiers in Oncology*, 10: 523543.

Chhoda, A.; Lu, L.; Clerkin, B. M.; Risch, H.; and Farrell, J. J. 2019. Current approaches to pancreatic cancer screening. *The American journal of pathology*, 189(1): 22–35.

Dutch Pancreatic Cancer Group (DPCG). 2023. Criteria Resectabiliteit. Accessed January 11, 2023.

Giannone, F.; Capretti, G.; Hilal, M. A.; Boggi, U.; Campra, D.; Cappelli, C.; Casadei, R.; De Luca, R.; Falconi, M.; Giannotti, G.; et al. 2021. Resectability of Pancreatic Cancer Is in the Eye of the Observer: A Multicenter, Blinded, Prospective Assessment of Interobserver Agreement on NCCN Resectability Status Criteria. *Annals of Surgery Open*, 2(3): e087.

Joshi, S.; Parbhoo, S.; and Doshi-Velez, F. 2021. Learning-to-defer for sequential medical decision-making under uncertainty. *arXiv preprint arXiv:2109.06312*.

Kwon, Y.; Won, J.-H.; Kim, B. J.; and Paik, M. C. 2018. Uncertainty quantification using bayesian neural networks in classification: Application to ischemic stroke lesion segmentation. In *Medical Imaging with Deep Learning*.

Lao, Y.; David, J.; Fan, Z.; Bian, S.; Shiu, A.; Chang, E. L.; Sheng, K.; Yang, W.; and Tuli, R. 2020. Quantifying vascular invasion in pancreatic cancer—a contrast CT based method for surgical resectability evaluation. *Physics in Medicine & Biology*, 65(10): 105012.

Lee, T. C.; Shah, N. U.; Haack, A.; and Baxter, S. L. 2020. Clinical implementation of predictive models embedded within electronic health record systems: a systematic review. In *Informatics*, volume 7, 25. MDPI.

Li, D.; Hu, B.; Zhou, Y.; Wan, T.; and Si, X. 2018. Impact of tumor size on survival of patients with resected pancreatic ductal adenocarcinoma: a systematic review and meta-analysis. *BMC cancer*, 18(1): 1–8.

Liu, J.; Gallego, B.; and Barbieri, S. 2022. Incorporating uncertainty in learning to defer algorithms for safe computer-aided diagnosis. *Scientific reports*, 12(1): 1762.

Marchegiani, G.; Andrianello, S.; Malleo, G.; De Gregorio, L.; Scarpa, A.; Mino-Kenudson, M.; Maggino, L.; Ferrone, C. R.; Lillemoe, K. D.; Bassi, C.; et al. 2017. Does size matter in pancreatic cancer? *Annals of surgery*, 266(1): 142–148.

Rigiroli, F.; Hoye, J.; Lerebours, R.; Lafata, K. J.; Li, C.; Meyer, M.; Lyu, P.; Ding, Y.; Schwartz, F. R.; Mettu, N. B.; et al. 2021. CT radiomic features of superior mesenteric artery involvement in pancreatic ductal adenocarcinoma: a pilot study. *Radiology*, 301(3): 610–622.

Taghizadeh, L.; Karimi, A.; and Heitzinger, C. 2020. Uncertainty quantification in epidemiological models for the COVID-19 pandemic. *Computers in Biology and Medicine*, 125: 104011.

Toshima, F.; Inoue, D.; Yoshida, K.; Izumozaki, A.; Yoneda, N.; Minehiro, K.; and Gabata, T. 2022. CT-diagnosed extra-pancreatic extension of pancreatic ductal adenocarcinoma is a more reliable prognostic factor for survival than pathology-diagnosed extension. *European Radiology*, 32: 22–33.

Zhang, J. 2021. Modern Monte Carlo methods for efficient uncertainty quantification and propagation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(5): e1539.