We highly appreciate the valuable feedback and suggestions of the editorial team and all reviewers. We have revised the manuscript (marked in red) accordingly and provided the point-to-point responses (blue) as follows to help address the reviewers' comments and suggestions.

Comments from the Editor

We agree with the reviewers that the performance of the model in a representative population, such as the ones you use in validation analyses, should be further characterized, with additional interpretability analyses.

Thank you for your valuable suggestions. In the revised manuscript, we extensively provided additional characteristics of the data that were required from the reviewers. Examples include additional patient and CT characteristics (Extended Data Table A1; Supplementary Table B5, B7), standard of truth determination (Section 5.2), data annotation (Section 5.2.2), data involved in model evolution (Section 5.3.5), and information of multiple scenarios in the real-world evaluation (Section 5.2.6).

The interpretability of the model was now analyzed in details. The performance of the interpretable segmentation maps was provided and discussed showing that our model can indeed locate the lesions well and precisely. In addition, we visualized the Grad-CAM heatmaps of PANDA Stage-2, to see which part of the image contributes most to the classification of abnormality. We also illustrated the top activated attention maps of the Transformer branch of PANDA Stage-3 to interpret how PANDA classified the lesions. The memory tokens of the Transformer not only attended the lesion locations but also considered the secondary signs for lesion diagnosis as utilized by the radiologists. The highlight of the interpretability analysis has been added to Section 2.2 (lines 273-280), Section 5.7 (lines 1283-1306), Supplementary Table B8, and in Extended Data Fig. A4 in the revised manuscript.

Please further refer to the following point-to-point response for details.

Comments from the Reviewer #1

The manuscript by Cao and colleagues presents a large scale analysis of noncontrast CT and AI for pancreatic cancer detection. The topic is timely and of high clinical relevance. Pancreatic cancer has a dismal prognosis and is predicted to be the second leading cause of cancer related mortality within the next decade. Therefore, novel options for early detection and/or screening of pancreatic cancer are urgently needed. The proposed AI based model achieves very high detection accuracy (sensitivity 92.9% and specificity 99.9%). The results are rather exceptionally and -if confirmed- could significantly alter our approach to screening for pancreatic cancer.

Response: We greatly appreciate your supportive comments and encouragement. In particular, the 99.9% specificity (i.e., one false positive in 1,000 image exams) is a highlight of our

proposed AI system, which will only produce 1/10 false positives than a system with 99% specificity, offering a promising prospect on minimizing false positives (and making the positive prediction rate clinically acceptable) to facilitate the future/next large-scale opportunistic screening of pancreatic cancer.

There are some aspects and points of criticism, which the authors should address:

1. One drawback of the present analysis is that the training set is from one center only. There is a risk of bias with this approach that should be discussed.

Response: Thanks for the suggestion. We have added discussions about PANDA's generalization in the Discussion session (lines 539-561) in the revised manuscript. Briefly, the mentioned risk of the PANDA model indeed exists but has been largely mitigated by multi-center validation and the fact that the internal training center, as a top tertiary hospital with clinical expertise specialized in pancreatic cancer, accept and treat patients with pancreatic diseases from all different parts of China. In addition, our PANDA Plus is trained on a "continual learning" approach using multi-center data (Section 5.3.5 in the revised manuscript).

2.External international multicentre study: most patients were from China and Taiwan ROC, only less than 4% of the patient were from outside China/Taiwan. This is a drawback that should be discussed further. Real international patient cohorts would strengthen the study.

Response: Thanks for the suggestion. International multi-center collaboration is a major challenge. During the manuscript preparation and revision process, we attempted to establish contact with several US hospitals, but within the revision period, we are unable to achieve effective collaboration yet. This will be our future work. Nevertheless, the generalizability of our model on the data from Taiwan ROC and the Czech Republic have given strong positive evidence for international multi-center generalization. At the CT level, patient imaging characteristics are not expected to change drastically across different races. Following your suggestion, we have renamed this cohort as "External Multicenter Test Cohort" and added discussions about the limited international data in the Discussion session (lines 657-661) in the revised manuscript.

3.It would be important to differentiate main-duct from side branch IPMN; first, since they have distinct morphological properties, second, since they have a different malignant potential, and third, since they have a vastly different incidence. If possible, this should be done in a revised version.

Response: Thanks for the valuable suggestion. In the revision, we added an additional module to subsequently classify the subtype of IPMN, i.e., main/mixed-duct IPMN vs. branch-duct IPMN. To achieve this, we first labeled the subtype of IPMN in the internal (SIPD) training and testing cohort. The pathologist (15 years of experience in pancreatic pathology) and the radiologist (17

years of experience in pancreatic imaging) in our team reviewed the surgical pathology records and images (when necessary) to determine the ground-truth of IPMN subtypes, which led to 163 main/mixed-duct IPMN and 91branch-duct IPMN in the training cohort, and 11 main/mixed-duct IPMN and 11 branch-duct IPMN in the internal test cohort. For the IPMNs in the external validation cohorts (n=172), due to the difficulty of retrieving and re-evaluating pathology records/images, the ground-truth of IPMN subtype was based on radiology evaluation by the radiologist by reviewing the multi-phase contrast-enhanced CT images (*Sahani et al, Radiology 238, 560–569, (2006)*), which led to 82 main/mixed-duct IPMN and 90 branch-duct IPMN . These additional data characteristics can be found in (Section 5.2.1, 5.2.3, 5.2.4) and Extended Data Table A1 in the revised manuscript.

We kept our original Stage-3 model (eight-class differential diagnosis) unchanged and trained an IPMN subtype classifier in a cascaded fashion, and then reported the classifier's performance separately. This is because we hope to keep our lesion differential diagnosis directly comparable with the radiology report, which has incomplete information on IPMN subtype diagnosis. We trained the IPMN subtype classifier with the same network architecture of our Stage-3 model on the IPMN cases in the internal training cohort, as Stage-3's attention-integrated architecture is well-suited for capturing the context and morphological structure, which is immensely helpful for this task. We report the classification result of the task of main/mixed-duct IPMN vs. branch-duct IPMN, with main/mixed-duct-IPMN as the positive class. In the new internal differential diagnosis cohort (n=87, an effort to address Reviewer#3 Q4), our model achieves an AUC of 0.944 (95% CI 0.894-0.982) with a sensitivity of 94.1% (95% CI 87.2-100) and a specificity of 80.6% (95% CI 66.7-93.1). In the external test cohort (n=172), our model achieves an AUC of 0.915 (95% CI 0.864-0.958) with a sensitivity of 89.0% (95% CI 81.6-95.4) and a specificity of 81.1% (95% CI 72.8-89.2). When taking Stage-3's lesion differential diagnosis result into account, in its successfully classified IPMN cases (Figure 2f, n=71 internally and n=139 external), our IPMN subtype classifier achieved an AUC of 0.948 (0.892-0.988) with a sensitivity of 95.5% (95% CI 88.9-100) and a specificity of 77.8% (95% CI 60.6-92.3) internally; and achieved an AUC of 0.947 (0.905-0.977), a sensitivity of 94.1% (95% CI 87.5-98.7) and a specificity of 80.3% (95% CI 70.8-88.9) externally. These results have been added to Section 2.2 (line 252-255), Section 2.4 (line 368-370), and Supplementary Table B16 in the revised manuscript.

4.For small PDACs (diameter<2cm, T1 stage), the sensitivity for detection was 85.7%. Especially if used for screening, detection of small PDACs is paramount, since the prognosis even for those T1 tumors is bad.

Response: Yes, the detection of small PDACs is very crucial. The 85.7% sensitivity on T1 stage achieved internally could potentially due to the relatively small number of cases (n=13 for T1) in the internal test cohort. In the external validation cohorts with many more cases (n=283 for T1), the sensitivity rises to 92.2% illustrating the robust performance of our model detecting small PDACs. On the other hand, the relatively low sensitivity for the T1 stage is a sacrifice for a high specificity (>99%), as the prevalence of PDAC is low in the asymptomatic population; if used in

a designed screening scenario, e.g., screening in high-risk population, sensitivity can be increased by adjusting the model threshold accordingly. We have added some of these discussions in the revised manuscript (Discussion section, lines 649-653).

5.Reader studies: the inclusion of 11 readers is a rather modest number for rigorous analysis, especially considering that residents and non-specialists were also included. This part of the analysis might also be omitted.

Response: Thanks for your valued comment. For a more convincing evaluation and added strength for our reader studies, we invited additional readers. For the first reader study, the total number of readers now increased from 11 to 33, including 7 additional specialists (from 4 to 11), 6 additional residents (from 5 to 11), and 9 additional general radiologists (from 2 to 11, previously community radiologists). The updated information of the readers are displayed in Extended Table A2. The same analysis was performed again with these additionally involved readers and the results were presented in Section 2.3 (lines 296-326 in the revised manuscript. Due to the time limit of the revision period, the wash-out period between the original reading and AI-assisted reading was reduced to at least 1 month (previously 6 months) in the first reader study. In summary, the newly obtained results and observations are similar to our previous findings. Please refer to Section 2.3 for more details.

6.Second reader study: considering the sample size, how relevant is the outperformance of PANDA by 3.8% in sensitivity and 1.2% in specificity?

Response: Thanks for mentioning this aspect. For the second reader study, we now increased the number of specialists from 11 to 15 to increase the statistical power. PANDA (on noncontrast CT scans) outperformed the average performance of the specialists (using contrast-enhanced CT scans) by 2.9% (95% CI 0.1%--5.8%, P=0.087) in sensitivity and 2.1% (95% CI 1.4%--3.0%, P<0.0001) in specificity, for lesion detection; and a margin of 13.0% (95% CI 8.5%--17.8%, P<0.0001) in sensitivity and 0.5% (95% CI -0.7%--1.9%, P=0.68) in specificity, for PDAC identification. As the P value didn't indicate significance for the sensitivity of lesion detection and the specificity of PDAC identification, we did non-inferior test (at a 5% margin) for these metrics and found non-inferiority (P<0.0001 and P<0.0001 respectively). Although the lesion detection improvements are not large, PANDA has a large lead (13.0%) of sensitivity in identifying PDAC from 'other' (nonPDAC + normal) than the specialists. We have updated our manuscript accordingly (lines 327-339). Note that the AI model PANDA performed on noncontrast CT scans while human readers used routine contrast-enhanced CT scans.

7.Ground truth labels for normal controls were confirmed by a two year follow-up. Does that mean that all patients without pancreatic pathology were followed up after two years?

Response: Yes, but not all were followed up. The normal controls were selected based on the following approach:

a. First, we searched patients whose radiology report of abdominal contrast-enhanced CT had negative pancreatic findings.
b. Then, among these patients, we searched patients who had a record of at least 2-year follow-up and no information in their available clinical diagnosis indicated a pancreatic lesion.

We have revised our manuscript for better clarity (Section 5.2, lines 697-702).

8.What was the indication to perform the first CT scan and what was done at the follow-up visit? Clinical examination, blood tests, ultrasonography? Please specify.

Response: To our understanding, the reviewer is asking a follow-up question for the normal controls, i.e., Question 7.

Because the medical centers in our studies are all (or directly affiliated to) tertiary or general hospitals, the indication of the normal control patients to perform the first CT scan includes various purposes, such as abdominal pain, abnormal blood biomarkers, tumor (other than pancreatic tumor) diagnosis, etc. At the follow-up visit after at least two years, we determined a patient as normal if no information in their available clinical diagnosis (such as electronic medical record, blood test, and radiology report, etc.) indicating a pancreatic lesion. We have added these into the revised manuscript (Method 5.2 lines 697-706).

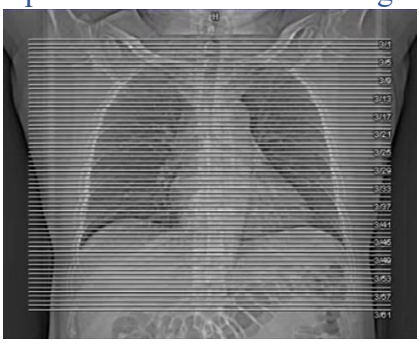
9.Model performance to chest CT: this is of interest, however, it largely depends on the chest CT protocol, i.e. how far in to the abdominal cavity the scan is carried out. Obviously this might differ from center to center and also nationally. Information regarding chest CT scan protocols should be provided.

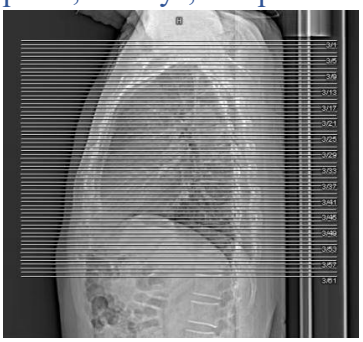
Response: Thanks for your encouragement. For the chest CT cohort, the protocol of chest CT is described as follows.

(1) Patient position
    a. supine position, thorax centered within the gantry
    b. both arms elevated.
(2) Respiration phase: single breath-hold: inspiration
(3) Scan extent: from the lung apices to the bottom

Note: Radiology technicians need to perform a CT localizer scan before determining the scanning range. Usually, relying solely on a single position's localizer image, such as the anterior-posterior view (Fig. A), may not accurately determine the scanning position of the lower lung border. Therefore, in our routine chest CT scanning protocol, a lateral view (Fig. B) is commonly added to better determine the scanning position of the lower

lung border. As a result, this often results in the scanning range of the chest CT covering a portion of the abdominal organs, such as the liver, spleen, kidneys, and pancreas.


(Fig. A)


(Fig. B)

(4) Tube voltage: 120 kV
(5) Tube current: 137-380 mA
(6) Scan geometry:
    a. FOV: 300mm – 418mm
    b. Slice thickness: 1mm / 5mm (we run PANDA on 5mm scan)
    c. Reconstruction kernel: lung kernel / soft tissue kernel (we run PANDA on soft-tissue-kernel scan)

We have added these descriptions into the revised manuscript (Section 5.2.5 lines 894-909).

10. Real world clinical evaluation: it is not clear what the different scenarios (physical examination, emergency, outpatient, and inpatient) actually are, i.e. what is the difference between physical examination and inpatient? This seems to be country or institution specific. Please clarify.

Response: You are correct that the specific terminology and definitions of different clinical scenarios may vary based on the country or institution. However, the general concepts behind these terms remain consistent. Our real-world clinical evaluation center is a tertiary hospital in Shanghai, China, which has all-inclusive medical services and benefit our multi-scenario real-world evaluation.

Physical examination center: This is indicated for routine check-ups, such as annual checkups. In this scenario, noncontrast CT scans, e.g., chest noncontrast CT scans for lung nodule screening or coronary artery calcium (CAC) scoring, are usually performed, which can be used for the opportunistic screening of pancreatic cancer. This scenario may be similar to the imaging centers in the U.S. medical systems.

Emergency department: Emergency room (ER) in the hospital, equipped with imaging devices (noncontrast CT is commonly used) to offer timely assessment of patients in acute diseases.

Outpatient: Outpatient care is defined as hospital care without being hospitalized or for a stay of less than 24 hours. Outpatient services encompass a wide range of diagnosis, treatment, or follow-up of various conditions.

Inpatient: Inpatient care refers to medical treatment provided to patients who are admitted to a hospital and require overnight stay or longer-term care. Inpatients typically have more severe illnesses and complex medical conditions. Inpatient care involves multidisciplinary teams of healthcare professionals working together to provide comprehensive management and treatment.

When considering the differences of these four scenarios in the application of opportunistic AI screening of pancreatic cancer, they have the following main differences in our study.

	Image background complexity	Contrast-enhanced (mostly)?	Pancreatic lesion (relative) prevalence	Radiologists' experience (first line)
Physical examination	Low	No	Low	Low
Emergency	Medium	No	Medium	High
Outpatient	High	Yes	High	Medium
Inpatient	High	Yes	High	Medium

Since the above differences and indications of the patients varied significantly among these four scenarios, separate evaluation on individual patient subpopulations is done to find out the feasibility of opportunistic screening using PANDA. These results could serve as an important reference when applied to different countries or institutions based on the source of patients. We have added key information of these into the revised manuscript (Section 5.2.6 lines 932-954).

11. It is stated that false positives by AI were for example common bile duct and pancreatic duct dilation. Since both might be a sign of an underlying pancreatic pathology, how was this excluded?

Response: We completely agree that both might be a sign of pathology, so we did not simply exclude these cases. The key point here is that these duct dilation cases do not have a lesion/tumor but might indicate underly pathology; therefore, simply determining their ground-truth label as 'positive' (having a lesion) or 'negative' (healthy) is not suitable.

As mentioned in our original manuscript (lines 424-439), we reported two results separately. (1) Include those patients with (peri-)pancreatic diseases (e.g., duct dilation) but without pancreatic lesions in the result calculation, and count the positive AI prediction as false positive; This is a rigorous calculation because AI picked up these disease-relevant signs, which should not be considered real false positives. (2) Exclude those patients from the result calculation, because they are clinically significant findings (Extended Data Fig. A6 in the original manuscript, e.g., a sign of underlying pancreatic pathology, as you mentioned); We did not count these patients as true positives because there is no evidence of pancreatic lesions in these patients. We have revised our manuscript for better clarity (lines 449-452 in the revised manuscript).

12. How was “pancreatic fatty infiltration” defined? Was this a radiological diagnosis on contrast enhanced CT, or MRI? Please specify.

Response: Focal fatty infiltration can mimic a hypoattenuating mass on CT scans, and fatty sparing of the pancreatic head can appear as a “pseudomass”. Previous studies (*Radiology 194, 453–458 (1995); Radiology, 190, 437–439 (1994); J. Comput. Assist Tomogr. 33, 90–95, (2009)*) have discussed the evaluation of fatty infiltration with CT imaging. For example, measuring CT attenuation on noncontrast or contrast-enhanced CT is a usually utilized method (*J. Comput. Assist Tomogr. 33, 90–95, (2009)*). In our work, we used the radiological diagnosis on contrast-enhanced CT (if available) or noncontrast CT (if contrast was not available). We have added some concise explanations about this in the revised manuscript (lines 477-478).

13. The real world test cohorts were from one center only, which is a drawback that should be discussed further.

Response: Thanks for your valued comment. We actually have performed another real-world study in Site B (FAHZU), including ~70,000 consecutive patients’ abdominal noncontrast CT. PANDA Plus maintained a similar high performance as in the current real-world center (SIPD). However, if we include this new large-scale real-world study in the current manuscript, we are concerned that the content will become excessive, as reviewer #2 has already noted that "the data and experiments are too extensive for a single work." Therefore, we are considering reporting this new real-world study in the next manuscript. Following your suggestion, we have added this when discussing the limitation in the Discussion section (lines 659-661).

14. Cost estimate in the discussion: it would not be necessary to perform MRI in each case; there would be other options such as CEUS, CE-CT or EUS.

Response: Thank you for your suggestion. As suggested, we evaluated the cost of performing CEUS, CE-CT, or EUS separately. We have revised the manuscript (lines 632-637) to report a range of cost estimates, considering these options could be used depending on different situations.

15. Screening would be important for high risk cohorts, e.g. age 50-75, newly diagnosed diabetes mellitus, obesity etc. The authors could model this from the real world population and present numbers needed to screen, sensitivity and specificity.

Response: Thanks for this interesting and potentially very valuable suggestion. Our real-world population currently only has limited records about newly diagnosed diabetes mellitus. Nevertheless, recent studies (as mentioned in the CAPS Consortium, *Gut 69, 1–17, (2020)*) have shown that 0.4% to 0.8% of patients with new-onset diabetes aged $\geqslant$50 will be diagnosed with pancreatic cancer within 3 years. Given the assumed sensitivity (93%) and specificity (99.9%) of PANDA for the task of PDAC identification in the retrospective real-world evaluation, null sensitivity and specificity of 50%, and a prevalence of pancreatic cancer among high-risk new-

onset diabetes subjects aged ≥50 years of 0.8%, we calculated the number of people needed to screen is 1,500 to achieve a statistical power of 90%. The calculation is based on Test for One-Sample Sensitivity and Specificity via PASS software (version 15). This result has been added to the Discussion section (lines 643-648).

Minor points:
1. Line 99: "approximately 466,003". If it is an approximation, state as 466,000.
2. Line 110: "Noncontrast CT is widely used". This is healthcare system dependent. It is not widely used everywhere.
3. Line 166: PNETs are neoplastic lesions that might require surgery (not usually, since small PNETs are commonly not an indication for surgery).

Response: Thanks for pointing these out. We have modified the respective contents accordingly (lines 103, 115, 175).

Comments from the Reviewer #2

A. Summary of the key results

The purpose of this study is to assess the performance of an automated pancreatic lesion detection and classification system based on a multilevel neural network design (PANDA model hereafter). The system operates on noncontrast clinical CT scans and generates a binary estimate of lesion presence, a categorical estimate of the type of lesion present, and spatial label maps of the pancreas and any identifier lesions.

The primary development cohort included 3208 subjects with abdominal CT scans that included a noncontrast phase, including 1431 PDAC cases, 938 controls, and 839 cases with other pancreatic lesions including a range of malignant, premalignant, and benign lesions plus chronic pancreatitis. This was assembled retrospectively using a case- control design. Classification was confirmed by surgical pathology for lesions or by two- year follow up for controls, although there are areas of uncertainty around the methods of case confirmation.

This study includes numerous internal and external validation experiments. The cohorts and experiments are as follows:

A. PANDA training set – n=3208 subjects with noncontrast CT scans, 1431 PDAC and 839 non-PDAC pancreatic lesions with surgical confirmation plus 938 controls without pancreatic lesions. Non-PDAC lesions included neuroendocrine tumors, solid pseudopapillary tumors, IPMNs, MCNs, serous cystic neoplasms, chronic pancreatitis, and 24 other benign and malignant entities.

B. PANDA internal testing set (n=291, with 108 PDAC, 116 normal, and 67 nonPDAC)
C. External multicenter test cohort (nine centers, n=5337, with 2737 PDAC, 932 nonPDAC, and 1668 normal)
D. Chest CT cohort (n=590, w 161 PDAC, 51 nonPDAC, and 378 normal)
E. "Real world" clinical test cohorts 1 (n=16420) and 2 (n=4110)

The study includes a complex array of experiments, briefly summarized as follows:

1. For binary lesion detection in the internal testing set, PANDA achieved an AUC of 0.996.
2. For lesion categorization in the internal testing set, PANDA achieved a total accuracy of 81.7% and a balanced accuracy of 65.1%.
3. For lesion categorization in the internal testing set by PANDA compared to a "second-reader radiology report", no significant difference was detected ($p>0.5$).
4. Ablation studies showed that PANDA Stage-2 had a statistically significant elevation in AUC versus nnUNet, albeit by a trivial margin (0.993 vs 0.988). PANDA Stage-3 outperformed a multitask CNN for classification (54.2% vs 46.7% balanced accuracy).
5. Reader studies showed poor performance of readers on both noncontrast and contrast CT scans. PANDA performance exceeded that of readers in both contexts. Joint reader/PANDA study showed increase in reader performance with AI prompts.
6. In the multicenter validation cohort study, PANDA achieved an AUC of 0.984 for binary lesion detection. For lesion categorization, PANDA overall accuracy was 81.4% with a balanced accuracy of 52.6%.
7. In the chest CT cohort, PANDA achieved an AUC of 0.961 for binary lesion detection.
8. In the real world cohort 1, PANDA achieved overall sensitivity of 84.6% and specificity of 99% for lesion detection with a positive predictive value of 56% for PDAC identification.
9. In Real World cohort 2, the "PANDA Plus" revised model achieved a sensitivity of 93% and specificity of 99% for lesion detection.

Response: We greatly appreciate your very comprehensive and thorough review of our work. And thanks for the nice summary of this work, we made supplementary explanations as follows and have revised our manuscript accordingly for better clarity.

Summary 4: The strong baseline method nnUNet achieved a high AUC value of 0.988, making further improvements challenging. In fact, we have tried several other approaches, such as radiomics + classifiers and nnUNet + CNN, with specific optimization effort. Only PANDA Stage-2 method was able to reliably improve the AUC value by 0.5% to reach 0.993. Notably, this improvement was statistically significant based on the evaluation of a large-scale dataset (n=3,208) with a p-value of 0.00022. Another perspective might reflect this significance better – at the same (desired) specificity level of 99.0%, PANDA Stage-2 outperformed nnUNet in sensitivity by 4.9% (95.2% vs. 90.3%). We have added a dotted line in Extended Data Fig. A2a to show this

improvement more clearly in the revised version, and revised the manuscript accordingly (lines 263-265).

Summary 5: In the second reader study (of the original manuscript), pancreas specialists showed good performance on contrast-enhanced CT with an average sensitivity of 91.1% and specificity of 98.8% for lesion detection. The best one achieved 95.4% and 98.3%.

Summary 9: In Read-world cohort 2, "PANDA Plus" showed an improved specificity than PANDA (99.9% vs. 99.2%). This indicated that the false positive rate decreased by about 85% (from 0.8% to 0.1%), achieving 1 false positive in 1,000 tests.

B. Originality and significance

B.1. There have been previous publications on the automated detection of pancreatic lesions and on the classification of pancreatic lesions, including the FELIX project cited within this paper. The specific AI architecture used in this study is novel and compelling.

Response: Thanks for recognizing our technical contribution.

C. Data & methodology: validity of approach, quality of data, quality of presentation

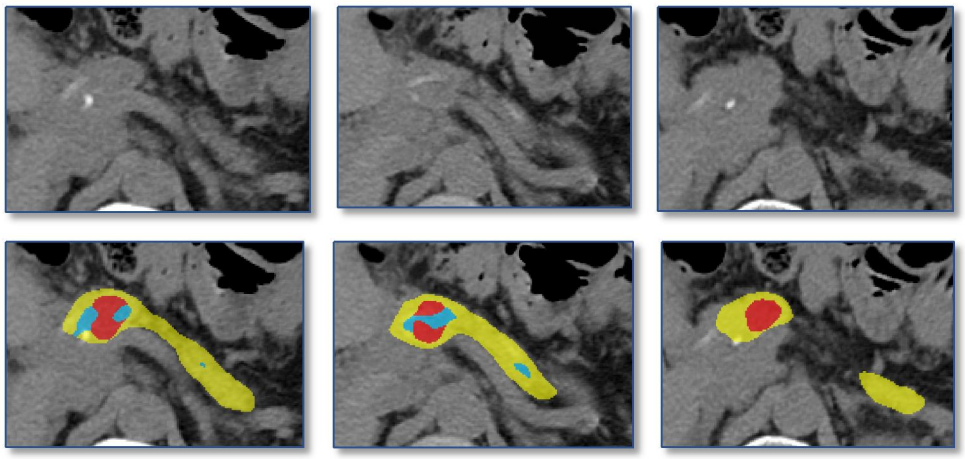
C.1. The design and methodology of the multilayer neural network system for localization, segmentation, and classification are state of the art and align well with best practices in the field. While the previous publications on the localization methods show Dice scores that are below the leading models in the literature, the subsequent use of attention models for classification is a reasonable strategy to overcome limitations in the semantic labels. The AI methods are described to an appropriate level of detail.

C.1.1 The novel components of the AI networks deserve detailed evaluation prior to the multiple cohort studies because they were trained de novo for this study. The network stages for pancreas segmentation and lesion segmentation should be presented with data on their pixel-level accuracy (e.g., Dice score and Hausdorf distance). The lesion classification model is adequately characterized in subsequent tables.

Response: Thank you for the suggestion. To evaluate the segmentation performance, we newly annotate the 291 cases in the internal test cohort following the method described in Section 9.2.2 in the original manuscript. The only difference is that the final pancreas mask on the noncontrast image was verified and edited by the radiologist (who annotated lesions) instead of the experienced engineer. Then, we evaluate the Dice score (DSC) and the 95 percentile of Hausdorff distance (HD95) for whole pancreas (including lesion area), healthy pancreas area and each type of lesion. Briefly, the DSC for whole pancreas, healthy pancreas area, PDAC, and nonPDAC are 0.903, 0.852, 0.719, and 0.703, respectively. The detailed results and the additional annotation process were added to the revised manuscript (lines 273-278, 787-788, and Supplementary Table B8).

C.1.2. Figure 4 – while the selected images depict compelling localization of lesions, the pancreas segmentations appear to bridge into adjacent fat. It would be helpful for the authors to provide data here on the segmentation accuracy (Dice score, Hausdorf distance) of the primary pancreas segmentation algorithm.

Response: Thank you for highlighting this. The pancreas segmentation accuracies of these three cases are 0.798, 0.886, and 0.901. The inaccurate feeling may be partially caused by the different cropped scales between the original CT image and the segmentation mask image shown in Fig. 4f, which could result in some visual bias. We have corrected this issue in the revised Fig .4f. Another factor could be the resampling process during the inference phase of the AI model. Specifically, the chest CT scans have a z-axis spacing of 5mm and were first resampled to 3mm (the median spacing of the training data) before model testing, following the standard data processing pipeline of nnUNet. Thus, the input image of the AI model will have smaller appearance gap between the slices than the original 5mm image. After testing, we resampled the AI generated segmentation mask to the original spacing (5mm) to produce the final segmentation result. The interpolation of the resampling process forth and back during inference will cause the output mask of the AI model to be smoother in the z-axis. This also applies to the x- and y-axis, potentially resulting in a slight decrease in segmentation accuracy at the boundaries. Here, we additionally visualize the upper and lower slice of the target slice and found that this is an intermediate slice between fat and pancreas body.



Original Slice shown in Fig 4 Upper Slice Lower Slice

In addition, thanks to your question, when double-checking this case, we found that it is the patient's chest CT after surgery – we have changed to his/her chest CT before surgery in the revised manuscript, where the re-calculated pancreas DSC for the prediction on the pre-operative CT is 0.844, and the PANDA prediction is PDAC (probability 89%). Moreover, we have re-checked all the patient cases in the chest CT test cohort to ensure all other cases were before surgery updated the related results accordingly (Fig. 4).

C.2. The descriptions of how the cohorts were assembled and how data abstraction was performed are superficial and lacking critical details. Several key issues are as follows:

C.2.1. The description of ground truth for lesions varies throughout the paper – "surgical pathology" in section 2, "pathology" in section 3 (line 219), "pathology diagnosis reports" in section 5 (line 328), "surgical pathology" in section 6 (line 364-365), "surgical pathology or biopsy" in section 9 (line 615), and "surgical or biopsy pathology" (line 711). Due to the wide range of the reported pathology, which includes a spectrum from benign chronic findings to aggressive malignancies, which of the lesions were confirmed only by surgical resection and which may have been confirmed by biopsy.

Response: We apologize for the confusion. The ground truth of the cases in the internal training, internal test, and chest CT test cohorts are all confirmed by surgical pathology. In the external and real-world cohorts, we provided the proportion of surgical pathology and biopsy pathology for each cohort and updated this information to Extended Table A1 and related parts in the manuscript.

C.2.1.1. Given that some of the findings are not typically biopsied or resected, including chronic pancreatitis and branch duct IPMNs without high risk features, the lack of specific consistent methods raises uncertainty about how the cases were curated and if significant selection bias towards high-risk lesions that warranted surgical intervention may be present. The authors should clarify and expand on these methods throughout the paper.

Response: Thanks for mentioning this potential bias. The cases in the internal training cohort and internal test cohort are all surgically resected lesions, including branch-duct IPMNs and chronic pancreatitis. This is because we aimed to build a definitive gold standard of lesion types for model learning. Surgical pathology is used as the inclusion criteria in similar studies (*Springer et al, Science Translational Medicine 11, eaav4772, (2019); Chu et al, Abdominal Radiology 47, 4139-4150, (2022)*). The cases in the external validation cohorts are either resected or biopsied. We agree that certain lesions, such as non-high-risk CP and branch duct IPMN, might not have sufficient number of or even without training data. However, this issue can be, to some extent, mitigated by the real-world cohort, where most pancreatic lesions are not resected or biopsied, but instead confirmed by standard-of-care clinical diagnosis. We mentioned this as the motivation of the real-world study in the original manuscript and have edited the text to enhance the clarity in the revised version (lines 404-409). To be specific, in the two real-world cohorts combined, only 5 out of 27 IPMN and 3 out of 97 CP cases were biopsied or resected (Table in responding your question

C.2.1.3; Supplementary Fig. B16-B22 in the original manuscript), and we achieved an overall sensitivity of 93% and 99% for detecting IPMN and CP, respectively (Fig. 5d), comparable with the performance on internal and external cohorts. This evaluation demonstrated that PANDA maintained robust performance in the real-world distribution of pancreatic lesions, although trained on surgical pathology-confirmed lesions. In our revised manuscript, we have included these in the Real-world section (lines 506-512).

C.2.1.2. Among the IPMNs, what was the distribution of main duct IPMNs, branch duct IPMNs, and mixed type IPMNs? What fraction of the IPMNs had high-risk stigmata or worrisome features (as per Fukuoka 2017 revised criteria or equivalent)? Among the BD-IPMNs, what were the indications for surgery?

Response: The distribution of IPMN subtype (main/mixed-duct IPMN and branch-duct IPMN) and high-risk/low-risk is updated in Extended Table A1. A pathologist in our team (15 years of experience in pancreatic pathology) and a radiologist in our team (17 years of experience in pancreatic imaging) reviewed the surgical pathology reports and slides (when subtype is not recorded) and contrast-enhanced CT to jointly determine the ground-truth, which led to 163 main/mixed-duct IPMN and 91 branch-duct IPMN in the training cohort and 11 main/mixed-duct IPMN and 11 branch-duct IPMN in the internal test cohort. For the IPMNs in the external test cohorts (n=172), the ground-truth of IPMN subtype was based on radiology evaluation, due to the difficulty of retrieving and re-evaluating pathology records. A radiologist in our team (17 years of experience in pancreatic imaging) annotated 82 main/mixed-duct IPMN and 90 branch-duct IPMN in the external test cohorts by reviewing the multi-phase contrast-enhanced CT images. The numbers are also displayed in the table below for the ease of reference.

Per Fukuoka 2017 revised criteria, the same radiologist reviewed the multi-phase contrast-enhanced CT of patients with IPMN in the internal training, internal test, and external test cohorts. The numbers of cases that have high-risk stigma, worrisome features (if no high-risk stigma is observed), and neither are observed are listed in the table below.

We have included the above information in our revised manuscript (Section 5.2.1, 5.2.3, 5.2.4).

Among the BD-IPMNs, the indication of surgery at SIPD center is one of the following being observed:

- growth rate ⩾5 mm per year
- increased levels of serum CA19-9
- acute pancreatitis (caused by IPMN)
- cyst diameter ⩾50 mm
- jaundice (tumor-related)

We have included these in our revised manuscript (Section 5.2.1, lines 729-

737; Section 5.2.3, lines 813-816; Section 5.2.4, lines 841-843).

No. of IPMN	Internal Train (n=254)	Internal Test (n=22)	External Test (n=172)
Main/mixed-duct IPMN	163	11	82
Branch-duct IPMN	91	11	90
High-risk stigma	154	10	101
Worrisome features (if no high-risk stigma observed)	94	11	70
Neither observed	6	1	1

C.2.1.3. The authors should provide counts of surgically resected versus biopsy confirmed lesions for each of the lesion categories in each of the cohorts.

Response: Thanks for the suggestion. For each of these cohorts, the ratio of the surgical pathology versus biopsy was updated in Extended Table A1. The internal training and testing cohort, Site A, D, E, F, G, H, and I are either all surgically resected, or only contained PDAC lesions. For Site B and Site C, we listed the count of surgically resected and biopsy confirmed lesions for each lesion category in the Supplementary Table B5 in the revised manuscript. For the chest CT cohort, all lesions were confirmed by surgical pathology. For the real-world cohorts, the required information is also displayed in the Supplementary Table B5 in the revised manuscript. And thanks to your suggestion, during the process of rechecking this information, we found that the original 161 PDAC in the chest CT cohort were actually the whole set of cases with pancreatic surgical pathology records, including not only PDAC (n=63) but also peri-pancreatic carcinoma and nonPDAC cases. We have corrected this issue and updated our results accordingly (Section 2.5, Fig. 4b,c,d).

C.2.1.4. What specific findings were required on pathological review to be classified into each of the groups? For example, high-grade dysplasia (also known as PanIN-3) has often been counted within PDAC statistics, as in the CAPS studies. How did the authors handle biopsies “suspicious for neoplasia” or “suspicious for adenocarcinoma” without a definitive diagnosis? Did the authors treat “adenocarcinoma with neuroendocrine features” as PDAC or as neuroendocrine tumors? This level of detail is crucial to understanding the validity of the classification results.

Response: According to the 2019 WHO classification of tumors (5th edition) of the digestive system, PanlN-3 was classified as a pre-cancer lesion, so we did not include them in our study. For biopsies without a

definitive diagnosis, we did not include such cases. Definitive evidence includes, for example, when malignant tumor cells are clearly observed; or in cases of uncertainty, immunohistochemistry can help determine malignant tumor cells, such as CEA, CA125, mesothelin, S100P, SMAD4, and p53 can be helpful in the distinction between invasive ductal adenocarcinoma and reactive pancreatic ductules. As for "adenocarcinoma with neuroendocrine features", adenocarcinoma containing scattered endocrine cells or enlarged residual nonneoplastic islets should not be mistaken for MiNEN (Mixed neuroendocrine—non-neuroendocrine neoplasms). Also, immunostaining for CK19 and MUC1/MUC5 can be used for confirmation of the ductal adenocarcinoma component. We include such cases as PDAC. No Mixed neoplasms in our internal training cohort or the multi-center validation cohorts. In the real- world cohorts, two cases of “mixed PDAC-PNET" were included (Supplementary Fig.B17 in the original version) and were categorized as PDAC for statistical analysis. Following your suggestion, we have added a paragraph to detail our pathological classification (Section 5.2 lines 692-696).

C.2.1.5. In the real world data study, line 463, the authors state that 89% of the false negatives were benign cysts, most under 10mm. They also state that surgical pathology was used to confirm cases. It would not be typical practice to resect small cysts without high-risk features. The authors should clarify that these were, in fact, resected or correct their selection methodology.

Response: Specifically, 53 benign cysts were included in real-world cohort 1, among which 39 were confirmed by standard-of-care diagnosis (SOC), and 14 were confirmed by multi-disciplinary team review (MDT). 11 benign cysts were included in real-world cohort 2, among which 6 were confirmed by SOC, and 5 were confirmed by MDT (Supplementary Fig. B16). A brief summary of this has been included in the Supplementary Table B5 (row 'Other’) in the revised manuscript by taking into account the comments from your review (C.2.1.3).

In our original manuscript, the standard of truth determination process in the real-world cohorts was described in Section 9.2.6 (lines 771-783) and in Extended Data Figure A5. The details of how each lesion was confirmed in each scenario, either by surgical pathology, SOC, or MDT, were reported in Supplementary Figure B15-B22.

C.2.2. For the control subjects in the internal cohort (page 6 line 176) and those recruited at specialist sites in the other cohorts, why were these subjects seen at a pancreatic disease referral center with a normal pancreas? What was the indication for a multiphase CT scan in these individuals?

Response: Because the centers in our studies are all tertiary hospitals, or pancreatic centers affiliated with tertiary hospitals (e.g., SIPD), they can access the multiphase CT scans of the patients performed for various indications, such

as abdominal pain and cancer (other than pancreatic cancer) diagnosis. We have revised related texts for enhanced clarity (Method 5.2 lines 702-706).

C.2.3. The remarkably high diagnostic accuracy of the AI models in this paper will require more detailed disclosures of potential confounders that may explain the performance of the methods. The authors need to convincingly show that model performance could not be due to extraneous post-diagnosis factors that would not be available in a prospective clinical environment, such as consistently different slice thicknesses, use of different imaging protocols for cases versus controls, presence of treatment-related hardware on case studies, differing scanner-specific reconstruction techniques that produce distinctive noise signatures, and obvious sequelae of clinical interventions like biliary stenting. For each of the cohorts, the authors should disclose the following separately for cases and controls:

- CT exam characteristics, including slice thickness, voxel size, z-axis extent, and median CT dose index (CTDI)
- Presence or absence of positive enteric contrast material (e.g., barium or gastrografin)
- Presence or absence of negative enteric contrast (e.g., oral water for pancreas protocol CT scans)
- Presence of biliary stents
- Were CTs acquired using a standard protocol for cases and controls in all cohorts? The multicenter study suggests not in an editorial comment in lines 351-355, but details are lacking elsewhere. Details and variations should be specified.

Response: Thank you for your valued comment. We first explain what efforts have been made to train a highly generalizable model while avoiding the potential result bias caused by confounders, and then provide the data characteristics following your suggestions.

By taking your and Reviewer #3's comments into account, we attribute our model's robust generalization ability as following, for more details please refer to our response to Reviewer #3's questions Q1 and Q2.

- Diverse training population
- Noncontrast CT is more generalizable than contrast CT
- Joint segmentation and classification model design improves generalization
- Threshold adjustment to produce a 99% specificity
- Model evolution increases the specificity to 99.9%

In addition, importantly, in terms of training data (please refer to the updated Table A1), the case and controls have similar CT imaging protocols (e.g., slice thickness, CTDI, oral water), forcing the model to focus on the primary learning objectives rather than fitting to shortcuts or confounders (e.g., different CT parameters, presence or absence of water/stents). In the testing phase, the model will thus have an "unbiased judgement" and achieved a sensitivity of 92.9% and a specificity of 99.9% in the real-world multi-scenario validation,

covering various imaging protocols and variations. We have involved the above analysis and summarizations in the revised manuscript (Discussion section, line 539-561).

We provided the required information as follows (updated in Table A1).

a. Please refer to Table A1 for more detailed information.
b. Positive enteric contrast material has not been used in any of the cohorts.
c. For the cases with pancreatic lesion, oral water was used in all centers except Site F and Site I. For the normal controls, oral water was used depending on the purpose of the abdominal CT scans. Please refer to Table A1 for detailed information.
d. All cases with biliary stents were excluded from this study.
e. Not exactly. Please refer to Table A1 for the statistics of the imaging protocols of each center. For cases and controls in the internal and external cohorts, although they were all noncontrast CT covering the pancreas region, they could be different, depending on the indication of abdominal CT scan (such as pancreas, liver, and adrenal gland examination) and each center's specific protocols. The chest CT cohort was chest CT protocol. The real-world cohorts were all protocols of noncontrast CT scans performed at SIPD.

C.2.4. Lesion size is a critical factor in lesion detection efforts but has not been reported in this work. What is the distribution of sizes of the lesions for each of the categories in the case group? What fraction of the lesions is less than 3cm in size? The authors do describe some analyses for small PDACs <2cm, but the overall distribution of sizes is unclear.

Response: Thank you for your suggestion. The distribution of lesion sizes and fraction of lesions less than 3cm have been added to the revised manuscript (Supplementary Table B6). The information of lesion size was either collected in the original surgical pathology report (if size was recorded) or was measured on the contrast-enhanced CT image.

C.2.5. Figure 4 – The chest CT images are labeled as "COVID-19 prevention Chest CT", but the methods state that the studies were acquired at SIPD, a pancreas disease center. The authors should clarify how the chest CT cohort was selected and why the "normal" subjects were being seen at a pancreas disease center.

Response: As our response to your question C.2.2, SIPD is a center affiliated with a tertiary hospital, so we can access the CT scans performed for various indications in the hospital. The collection of the chest CT cohort was described in Method section 9.2.5 in our original manuscript, and we have revised it to enhance clarity (now is Section 5.2.5 in the revised manuscript).

C.2.5.1. For the chest CT cohort (9.2.5), the provided range of "-20-307 days" (line 758) suggests that some of the studies may have occurred after surgical resection because the range spans positive and negative values around zero. The authors should clarify if any postoperative chest CTs were used in this cohort and should strongly consider excluding those cases from the analysis if present.

Response: We apologize for the oversight. The chest CT should be range of -20 – 307 days from the contrast-enhanced abdominal CT diagnosis. In other words, the chest CTs are either before or after the contrast CT diagnosis, and are all obtained before surgery. We have corrected the related descriptions (Section 5.2.5, lines 889-891). Note that in an effort to address your Question C.2.1.3., the range has been changed to "-20 – 191 days" since some data were excluded.

C.3. Human annotations were performed on contrast-enhanced CTs, not noncontrast CTs, and transferred to noncontrast studies through deformable registration. While defensible, this methodology is unusual and does raise concern about the accuracy of the ground truth pixel labels.

Response: Thanks for pointing this out. This is, in fact, the key technical "trick" or novelty of this work (lines 521-526 in the original manuscript). Radiologists find it extremely difficult to annotate on noncontrast CT images as lesion boundaries are almost invisible and hard to define (especially for PDAC), even when referring to contrast-enhanced CT images. To assure the reliability of the annotation, the final lesion mask of each case was verified and edited by an experienced radiologist to avoid obvious registration displacements. In fact, DEEDS performs reliably well in pancreas registration task, especially compared to other non-solid organs with larger deformations. Another benefit of using registration technique as a starting point is that the size of the annotated lesion can be consistent to the one appeared in the contrast-enhanced CT. We have added these comments in the revised Method Section 5.2.2 (lines 763-769). This is indeed an advantage.

C.4. There are several likely biases inherent to this retrospective case control design that should be acknowledged:

C.4.1. Exclusion of patients with acute pancreatitis decreases generalizability as this is a common confounder of PDAC.

Response: We agree that the ability to recognize acute pancreatitis is important for a full-spectrum pancreatic lesion detector, especially for real-world "bedside" clinical use. Therefore, after the first real-world study, to answer our question (5), we upgraded the model to obtain the PANDA Plus (Fig. B23 in the original manuscript), which had the ability to work on patients with acute pancreatitis. As shown in the original Fig. 5d, 90% of the patients with acute pancreatitis were detected, and all of the detected patients were correctly categorized into the nonPDAC class. We have revised the manuscript to improve clarity (Section 3,

lines 498-501).

C.4.2. Cohort selection bias is likely because cases are all clinically detected lesions. Real-world performance is likely overstated because the difficult-to-identify lesions would not have been included.

Response: In the real-world cohort the risk may have been reduced by the process described in ``Reference Standard" in Extended Figure A5, where 1% of cases that were found negative by both the standard of truth determination procedure and the AI prediction were reviewed by the two radiologists to confirm negative. No additional positive lesion was found in this process, showing that the impact of underlying difficult lesions was limited. In addition, some difficult-to-identify lesions should be in our chest CT and real-world cohort. For example, the cases shown in Fig. 4f were all missed by radiologists on the chest CT scans (there were quite a few such cases in that cohort, especially for the patient who had PDAC in the chest CT 307 days before the diagnosis and was missed by the radiologists at that time), and the 31 cases presented in Table A3 were missed by the standard-of-care (SOC). Our model could successfully detect these examples.

C.5. The data for the real-world validation cohort 1 are too sparse for a complete evaluation. The authors appear to reclassify half of the false positives (line 430), subsequently reporting the lower number (line 458). The authors should report the total false positive rate without modification. It is reasonable to provide subgroup analyses for the fraction that they feel to be clinically relevant findings that are not primary pancreatic neoplasms.

Response: As noted by Reviewer #1's Question 11, some false positives might be a sign of pathology, so we did not simply include or exclude these cases. Briefly, we reported two sets of results separately: (1) the clinically relevant findings were counted as false positives and (2) the clinically relevant findings were excluded from the result calculation. We have revised our manuscript for better clarity (lines 449-452).

The total specificity without modification is 99% for lesion detection (line 426 and Fig. 5b in the original manuscript); and is 99.8% for PDAC identification (line 427 and Fig. 5c). The stratification analysis of the categories of these findings was shown in Fig. 5h and Fig. A6 in the original manuscript.

C.6. Table A3 – This table shows that only 2 of the 26 cases that were positive in RW1 were actionable pancreatic neoplasms (cases 1 and 16). For all of these cases, it would be helpful to provide size measurements for the identified lesion. It would also be helpful to provide a similar table for the false positive results in RW1/RW2 to understand the likely tradeoffs in a real world application.

Response: The size measurement of the lesion was provided in Table A3 column "Maximum diameter at initial SOC / follow-up SOC (mm)" in the original manuscript. The false positive results were reported in our original manuscript (lines

458-462, 477-479). Following your suggestion, we have provided a table for false positives for RW1 and RW2 in the revised manuscript (Supplementary Table B17).

C.7. A number of figures provide marginal added value in this already-extensive manuscript and should be considered for removal or consolidation:

C.7.1. Figure A6 – This figure provides anecdotal descriptions of some of the false positives from the PANDA model. It adds relatively little value.

C.7.2. Figure A7 – This figure contains editorializing language that should be removed: "PANDA changes SOC diagnosis and Benefits patient". This figure is dedicated to a single case within the study and should be removed for brevity.

C.7.3. Figure B14 – This figure does not add to the science in this manuscript.

C.7.4. Figures B15-B18 and B19-B22 should be combined into one figure per study (RW1 vs RW2)

C.7.5. Figure B23 – This figure does not add to the science in this manuscript.

Response: We agree that this manuscript contains extensive content, particularly due to the addition of the "real-world" section, which is a notable contribution compared to most previous medical AI papers. The five figures mentioned by the reviewer are all from this section. As you mentioned earlier, biases exist in models developed in laboratory settings, no matter how carefully designed and validated across multi-centers. Therefore, these models are likely to fall short in the complex and variable real-world. Therefore, large-scale validation and analysis in real-world settings, as well as iterative model upgrades, are crucial for closing the clinical translation gap (*Rajpurkar, et al, N Engl J Med 388, 1981-1990, (2023); Lohmann, et al, Lancet Digit Health 4, e841-e849, (2022)*). Real-world validation introduces a series of intriguing new questions, and we try to provide informative and concise display items to address your comments.

We note that Nature journals allow for the inclusion of up to 10 items of Extended Data Figures/Tables after the main text (corresponding to Figure A6 and A7) and an independent Supplementary information document of unrestricted length (corresponding to Figure B14, B15, and B23). We have made revisions below.

(1) Figure A6 – following your suggestion, we have moved this figure to Supplementary information. These examples of (peri-)pancreatic disease findings shown in this figure were confirmed by either standard-of-care diagnosis or multi-disciplinary review with the reference of contrast-enhanced CT or MRI if available, which made these findings reliable. These examples can help healthcare providers/users gain a comprehensive understanding of the performance of PANDA in real-world clinical applications – some of its findings may not be true positive pancreatic lesions, but they are not negligible false positives either.

(2) Figure A7 – this is a valuable case to show the reader that the new AI model brings real patient benefits after clinical deployment. We note that very few imaging AI studies reported a real impact on the intervention of a patient with cancer. We highlighted this case in several parts throughout this main text, such as the Introduction section ("in some cases enabling timely treatment with intent to cure"), Section 7 (lines 452-457; "It is possible to detect malignancies ... and benefit …

cancer treatment"), and Discussion section ("already cured one patient with PNET"). Following your suggestion, we have changed the language to "PANDA screen-detected PNET".

(3) Figure B14 – this figure is to provide a solution for large-scale, real-world study to boost feasibility and efficiency. For large-scale studies involving CT imaging, it is very time-consuming and tedious to perform such studies in the traditional pipeline, i.e., downloading 20,000 patient data one-by-one manually from PACS, data anonymization, data curation, and running model in developing environment. As shown in our solution, we provide an engineering solution that seamlessly integrated the AI system into existing clinical infrastructures. This solution could facilitate future large-scale clinical studies/trials in hospital environment and thus we humbly ask to keep this figure in the Supplementary document.

(4) Figure B15-B18 and B19-B22 – following your suggestion, we have combined them into one figure per study (Fig. B16 and Fig. B21 in the revised Supplementary). We still kept the original ones, as readers may want to know the details for each clinical scenario. If the current format is unacceptable, we will remove the original ones.

(5) Figure B23 – this figure is one of the key contributions in our real-world study and answers our proposed question (5) Can the "benchtop" derived AI be further improved according to "bedside" clinical requirements? With the model evolution approach depicted in this figure, PANDA Plus significantly reduced the false positives by more than 80%, reaching a desired specificity of 99.9% for both pancreatic lesion detection and PDAC identification. We note that Reviewer #1 considered such results (sensitivity 92.9% and specificity 99.9%) to be rather exceptional and -if confirmed- could significantly alter our approach to screening for pancreatic cancer. Note that a screen test, even with a 99% specificity, is considered not appropriate for PDAC screening in the general population because the PPV is about 1% (*Grossberg, et al, CA Cancer J Clin 70, 375-403, (2020)*); while our 99.9% specificity corresponds to a PPV of 10%. In addition, recently, clinical researchers have begun to realize that post-deployment monitoring and upgrading are essential to ensure AI models' effectiveness and reliability in clinical settings (*Rajpurkar, et al, N Engl J Med 388, 1981-1990, (2023); Lohmann, et al, Lancet Digit Health 4, e841-e849, (2022)*). Taken together, we humbly ask to keep this figure in the Supplementary document.

C.8. Table B5 – Section 3 line 233 describes this data as derived from a "second-reader radiology report" while the table calls it "standard-of-care". It is unclear if this is a structured read with a directive to rank a differential diagnosis, an unstructured report from a study reader that was then abstracted into a ranked differential, or a secondary analysis a primary standard of care clinical radiology report. Greater detail should be provided in the methods and in the legends of this table and the associated figures.
Response: We apologize for the confusion. This is the third one you mention – a secondary analysis of a primary standard of care clinical radiology report, resulting from the double reading process. We added this detail to the manuscript in the

legends of this table and modified/edited the unclear description (now is Table B7 in the revised version).

C.9. Reader study (line 257-265) – Additional detail should be provided about the setup and conduct of the reader studies. Were radiologists allowed to use windowing tools in reviewing noncontrast exams? Was the review performed in a clinical PACS environment? What steps were taken to ensure that the readers were making a good faith effort to detect cases?

Response: Yes, windowing tools are available and allowed in the reviewing process. The review was not performed in a clinical PACS environment, because patients in the PACS system are with their clinical information and records, and it is not convenient to access the PACS system for readers from external centers. We tried to make sure the readers made a good faith effort via the following details.

(1) Each reader was trained prior to the study to properly use an independent CT viewing software (ITK-SNAP http://www.itksnap.org/pmwiki/pmwiki.php NIH supported). Basic functions of this software include but not limited to HU value windowing, zooming in and out, axial/sagittal/coronal view simultaneous display.

(2) The readers were informed that (i) they only needed to inspect the pancreas without time constraints, and (ii) the study dataset was enriched with more positive patients than the standard prevalence of pancreatic lesions in daily practice. This setup would benefit their sensitivity but may lower their specificity, compared to their routine work of reviewing CT images.

(3) The readers were recruited based on an interest in our AI system, especially in whether the AI system could perform better than the radiologists on this challenging task and whether this system could positively assist in improving the diagnostic level and accuracy of the radiologists.

Some of the details mentioned above have already existed in our original manuscript, and we have enriched the content thanks to your comments (lines 1233-1235).

C.10. Reader studies with contrast, figure B10 – The performance of the study radiologists on a set of cases with clinically established lesions is surprisingly poor, particularly among pancreatic specialists. The authors should discuss potential reasons why their readers achieved low sensitivity on this task. The authors should also provide a comparison between the study readers and the original clinical report that was generated with the clinical CT scan.

Response: Thank you for the suggestions. In the reader study on contrast CT, the readers achieved an average sensitivity of 92.0% and a specificity of 97.9% on lesion detection (our updated experiment after adding more readers to address Reviewer #1's question). This performance is rational for a study conducted using contrast-enhanced CT scans, especially considering that our dataset consists of a full spectrum of pancreatic lesions and early-stage pancreatic cancer. In comparison to recent clinical studies, the sensitivity of radiologists' performance conducted using contrast CT (either biphasic protocols or a single venous phase) to detect PDAC is 88.4%-100%

(*Park et al, Radiology 306, 140-149, (2023); LeBlanc et al., European Radiology, 1-9, (2023); Liu et al, Lancet Digital Health 2, e303-313, (2020)*) and to detect pancreatic lesions is 91.7%-95.6% (*Park et al, Radiology 306, 140-149, (2023)*), with a specificity of 93.2%-96.2% (*Park et al, Radiology 306, 140-149, (2023)*). As a reference, the best performing specialist in our study has a higher sensitivity of 97.1% and specificity of 99.1% in pancreatic lesion detection.

In terms of the radiology report, radiologists have complete access to the patient's clinical history (e.g., CT examination indicated for chronic pancreatitis follow-up), and the results of other clinical examinations (e.g., tumor biomarkers). The performance gap between the readers and the report indicated that this information was useful for the differential diagnosis of the lesions.

We also note that the lesion detection performance (e.g., 100% sensitivity) of the clinical report for the patients collected in our study cannot represent the real performance. If a case were not detected in the clinical report, the case would probably not have undergone resection, thus would not have been included.

According to the above information, we have modified the manuscript for better clarity (lines 575-578, lines 1182-1187).

C.11. Several of the authors are employed by a commercial entity (Alibaba). It is unclear from the manuscript if Alibaba holds commercial rights to the algorithms and software described in this paper. The editors should ensure that the manuscript complies with the journal's conflict of interest policies.
Response: Alibaba has filed for a patent disclosure for the work related to the methods of detection of pancreatic cancer in noncontrast CT. This is a common practice for companies, e.g., DeepMind (*De Fauw et al., Nature medicine 24, 1342-1350, (2018)*), Google Health (*McKinney et al, Nature 577, 89-94, (2020); Liu et al, Nature Medicine 26, 900-908, (2020)*). Authors have no other conflict of interest. We have disclosed this information to the editors in the submission questionnaires – competing interests policy.

D. Appropriate use of statistics and treatment of uncertainties

D.1. The results of the second-reader radiology study showed "comparable" performance with $p>0.5$ (line 233). This is not a formal test of equivalence. A formal test of equivalence like the two one-sided tests (TOST) approach should be provided.

D.2. Kappa statistics should be provided for classification accuracy to additionally capture how well the system performs beyond random chance in an imbalanced cohort.

D.3. For estimates of diagnostic performance, 95% confidence intervals should be provided (e.g., lines 425-429).

Response: Thank you for your suggestion. We have revised the statistical methods as follows.

(1) We tested the non-inferiority at a pre-specified 5% margin (*McKinney et al, Nature 577, 89-94, (2020)*), and found the accuracy on the internal differential diagnosis

cohort is non-inferior to the second-reader radiology report in accuracy (p=0.0009). We have revised accordingly (lines 246-249). Note that we have enriched the differential diagnosis cohort (n=768) in an effort to address Reviewer #3's Question Q4.

(2) Cohen's Kappa statistics of each center for the classification of differential diagnosis are shown below. These results have been added to the corresponding confusion matrix respectively (Fig. 2f, Supplementary Fig. B12).

	Internal Differential Diagnosis(n=786)	External Test (n=3669)	Site A (n=1274)	Site B (n=1506)	Site C (n=176)	Site D (n=254)
Kappa	0.66	0.59	0.52	0.63	0.49	0.65
Interpretation	Substantial	Moderate	Moderate	Substantial	Moderate	Substantial

(3) The confidence intervals have been added to the revised manuscript.

E. Conclusions: robustness, validity, reliability

E.1. The study currently suffers from uncertainties about its validity and robustness due to the absence of critical details in the cohort descriptions and management of confounding features. These may be remedied through improved reporting of study details.

Response: Thank you for your valuable comments and suggestions. In the revised version, we tried to provide sufficient critical details of this study per the reviewers' request. We hope the additional details will enhance the certainty and confidence of our work in terms of validity and robustness.

E.2. In chest CT cohort, the reported "detection" of lesions not in field of view through secondary signs highlights a limitation of this selected case-control design. For example, ductal dilatation is not specific to PDAC, but limited ductal dilatation seems to be present in the selected non-PDAC cases based on the authors discussion of this as a definitive finding by the algorithm.

Response: In the chest CT cohort, the number of the lesions that were not scanned of the nonPDAC group (n=6) is limited, due to the difficulty of case collection of chest CT from patients with pathology-confirmed pancreatic lesions. And yes, only limited ductal dilation was presented in the 6 cases (listed in the table below): 2 with ductal dilation were detected; 3 without ductal dilation were missed; the other one with only minimal ductal dilation was missed. This number does not suffice a critical number for the analysis of the ductal dilation detection on nonPDAC cases. We have added a discussion about this in the revised manuscript (lines 401-402).

Case No. (Lesion not scanned)	Pathology	PANDA detected	Note
06365632	SPT	No	Body, no ductal dilation observed
06436135	PNET	No	Body, no ductal dilation observed

06364512	IPMN	No	Head, only pancreatic tail scanned, minimal ductal dilation observed
06397732	CP	Yes	Head, ductal dilation observed
06363490	CP	Yes	Head, ductal dilation observed
06394426	CP	No	Head, only pancreatic tail scanned, no ductal dilation observed

E.3. Inclusion of simulated low-dose chest CTs when low-dose CTs were not available (383-390) – This section should be completely removed. Simulated data is not sufficient to establish efficacy and validity at this level. Speculation that simulated data shows validity for large-scale opportunistic screening is premature.
Response: Following your suggestion, we have completely removed the related contents.

F. Suggested improvements: experiments, data for possible revision

F.1. The data and experiments presented in this manuscript are too extensive for a single work. The result of this overloading of content is that critical details are either missing or are difficult to locate across the multitude of cohorts and experiments. It is frequently necessary to cross-reference two or three times to understand fundamental details about each cohort. The authors should consider breaking this work up into several manuscripts to allow for expanded details for each of the experiments and to improve the ability of readers to interpret and judge the contents. The current format is not tenable.
Response: We understand the reviewer's concern and sincerely appreciate the suggestion. In the previous sections of the response, we made every effort to provide the missing details requested by the reviewer and revise the manuscript accordingly to enhance the clarity whenever possible. We hope the revised version can adequately address the concern of missing details and readability.
Taking into account your comments and considering the current trend in medical imaging AI research, which typically includes sections on internal validation, external validation, and reader study, and considering the real-world validation is a distinctive feature of our current work (more explanations are as follows), we have restructured the manuscript. Please review the revised version.
As mentioned in recent review papers (*Rajpurkar et al., New England Journal of Medicine 388.21, 1981-1990 (2023); Lohmann, et al, Lancet Digit Health 4, e841-e849, (2022)*), "real-world evaluation" and "continual learning" are among the essential components of generalization safeguards for AI systems deployed in radiology and oncology. Our real-world study directly validated PANDA on real-world consecutive population of patients and upgraded the model ("PANDA Plus") on real-world feedback from the clinicians, i.e., false negatives, false positives, and acute pancreatitis, to address the feasibility and effectiveness of the aforementioned "continual learning". Results on real-world cohort 2 show that the AI model gains new abilities (e.g., detecting pancreatitis) after model evolution while retaining previously acquired knowledge. We hope that the integration of these contents will strengthen the

validation of PANDA as a generalized AI model in radiology.
In fact, if the multi-center validation and the real-world validation are viewed as a whole, several concerns raised in the reviewers could be addressed, for examples,

(1) Reviewer #2 – Question C.2.2.1., the concern that the cases (IPMN and CP) collected in the multi-center validation are biased towards the resected or biopsied. This concern could be addressed by the performance of our model on the real-world cohort where most IPMN and CP are clinically diagnosed (Extended Table A1).
(2) Reviewer #2 – Question H.1., inappropriate to make claims that this manuscript supports the use of PANDA for opportunistic screening based on the data that they present. The real-world study was designed to validate the AI model to leverage the incidental imaging data unrelated to the clinical indication, e.g., chest noncontrast CT, which conform to the definition of opportunistic screening in radiology (*Pickhardt, Radiology 303.2, 241-254(2022)*).
(3) Reviewer #3 – Questions Q1 and Q2, how do you make sure that AI models trained on a biased dataset generalize to the general population? By testing our model on the real-world consecutively collected patients with about 99% normal patients, we demonstrated that PANDA maintained high specificity (99.2%). We indeed encountered false positive conditions and with model evolution, we successfully boosted PANDA to a specificity of 99.9% on real-world cohort 2.

F.2. The authors present two versions of their AI model – PANDA and PANDA Plus – but data on the revised "Plus" model is limited (lines 465-485). They also introduced a new class, acute pancreatitis, into the model with minimal validation data (479). It would be better to separate the evaluations of these models into distinct papers.

Response: Thank you for your comment. "PANDA Plus" was trained on both the original training cohort and the newly collected "hard examples" and cases with acute pancreatitis (AP), not using the new data alone. We have revised the corresponding content in the manuscript (lines 488-489) to avoid confusion. The details of model upgrade, including the training details were originally described in Methods 9.3.5. In our revised version, we report more details of the training data of "PANDA Plus" (Section 5.3.5, lines 1132-1143).

The validation of PANDA Plus was performed on real-world cohort 2 (n=4,110), where PANDA Plus achieved 90% sensitivity on 40 patients with AP. We have revised the manuscript for improved clarity (lines 499-501). We plan to report the validation of PANDA Plus on even more real-world collected patient cases from FAHZU in future work. Please refer to our response to Reviewer #1's Question 13 and Reviewer #3's Question Q2.

F.3. Discussion, line 527 – The interpretable component of this system is highly compelling, but the authors have provided no primary data to validate that part of the system. The data from the reader study are presented with insufficient depth to assure validity of this component. The authors should provide additional validation data,

including Dice scores and Hausdorf distances, to validate that the interpretable probability maps are concordant with the imaging locations of the pathologically confirmed lesions.

Response: We appreciate your encouragement. The performance of the interpretable segmentation maps was provided in response to your question C.1.1. The segmentation performance shows that our model can locate the lesions well. Briefly, the DSC scores for segmenting the whole pancreas (including lesion area), healthy pancreas area, PDAC, and nonPDAC are 0.903, 0.852, 0.719, and 0.703, respectively. In addition, in an effort to address Reviewer #3's Question Q5, we visualized the Grad-CAM heatmap of PANDA Stage-2, to see which part of the image contributes most to the classification of abnormality. We also visualized the top activated attention maps of the Transformer branch of PANDA Stage-3 to interpret how PANDA classified the lesions. The memory tokens of the Transformer not only attended the lesion site locations but also considered the secondary signs for lesion diagnosis as utilized by the radiologists. The highlight of the analysis has been added to Section 2.2 (lines 273-278), Section 5.7 (lines 1283-1306), Supplementary Table B8, and in Extended Data Fig. A4 in the revised manuscript.

F.4. Section 9.6.1 – To determine the extent to which readers were unduly influenced by the AI data presentation, the authors should conduct an additional study in which false positive AI guidance prompts are introduced. If presented with lesion maps that are incorrect, how often do the radiologists incorrectly agree that a lesion is present? This will assist in understanding the impact of false positives in real world applications.

Response: Thank you for this interesting suggestion. We could not perform such an analysis as the AI did not produce false positives in the internal test cohort used for our reader study. In an effort to address your Question C.6, we provided a detailed categorization of the false positives in the real-world scenario (Supplementary Table B17). Over 90% of the false positives in the real-world cohorts are easy to rule out by the radiologists, and the remaining requires MDT's diagnosis and follow-up (lines 457-462, 477-479 in the original manuscript). Particularly, PANDA Plus's four false positives are all easy to be ruled out (lines 477-479 in the original manuscript). We hope this analysis can strengthen the validation of PANDA's safety in real-world applications.

F.5. Reader studies (line 257+) – The authors should provide a specific analysis of the individual differences in radiologist performance between the noncontrast and with-contrast CT scans. This analysis will provide useful context to the performance of the PANDA method on noncontrast studies.

Response: In the reader studies, we avoided the overlap of the readers between the noncontrast study and the contrast-enhanced study. Because we aimed to measure the individual performances of the readers either on noncontrast or contrast-enhanced CT, while the simultaneous examinations had the possibility of mutual interference.

Following your valued suggestion, we extended the study as follows. Four pancreas specialists additionally reviewed the noncontrast CT after a long wash-out period (about one year), and we evaluated the performance comparison below. Note that we did not tell these specialists about their results before this experiment. It can be observed that almost all readers' performance (sensitivity and specificity) in contrast-enhanced CT is superior to their performance in noncontrast CT. In addition, their average performance on the noncontrast CT is similar to the average of the other 11 specialists in the first reader study (82.0%, 96.9%; Table B9(a) in the revised version), and interestingly, their average performance on contrast-enhanced CT is similar to the average of the other 11 specialists with AI assistance on noncontrast CT (89.5%, 98.7%; Table B10(a) in the revised version). We have added this experiment in the revised manuscript (Section 5.6.2, lines 1268-1282).

	Noncontrast CT		Contrast-enhanced CT	
Specialist ID (ID in original manuscript)	Sensitivity	Specificity	Sensitivity	Specificity
S12 (S5)	84.6	98.3	92.0	98.3
S14 (S7)	80.0	99.1	85.7	99.1
S17 (S10)	93.7	88.8	95.4	98.3
S21 (S14)	82.3	100	85.7	98.3
Mean	85.1	96.6	89.7	98.5

G. References: appropriate credit to previous work?

Yes.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

H.1. Early detection versus automated detection – The authors make several statements suggesting that their data shows that this system will be useful for early detection of pancreatic cancer. While this is an important goal that we are all eager to achieve, it is critical that the manuscript accurately portray this work as being conducted in scans acquired at or after diagnosis. The authors have not suggested that any prediagnosis imaging – studies occurring at least 3 to 6 months prior to the clinical diagnosis and lacking an actionable finding that led to the diagnosis – were included in this study. This cohort does not provide insight into the early detection capabilities of this system, at least for PDAC.

This is a critical distinction in PDAC, where even regular annual surveillance strategies (e.g., the US and international CAPS studies) have managed to detect less than half of incident cancers in high-risk populations. Further data will be needed to show that detection of diagnosable lesions will translate into improved outcomes in

the general population.

The authors should limit their discussions of early detection and screening throughout the manuscript to speculation about potential future applications within the discussion section. Representative examples of inappropriate discussions of screening are present in the abstract (page 2 line 56), introduction (page 4 line 122 and 129), and Opportunistic Screening using Chest CT (page 11 line 356). It is specifically inappropriate for the authors to make claims that this manuscript supports the use of PANDA for opportunistic screening based on the data that they present because there were no specific prediagnosis case images included in the study.

Response: Thank you for the valuable suggestion. In our study, we follow the concept, efforts and the definition of early detection of PDAC in recent articles (*Dbouk et al., Journal of Clinical Oncology 40, 3257-3266, (2022); Placido et al., Nature Medicine 29, 1113-1122, (2023); Pereira et al., Lancet Gastroenterology & Hepatology 5, 698-710, (2020); Singhi et al., Gastroenterology 156, 2024-2040, (2019); CAPS Consortium, Gut 69, 1–17, (2020)*). For example, the JCO paper states, "early detection, at a stage when the disease is most curable (stage I PDAC), or when there is only a noninvasive neoplasm with high-grade dysplasia (HGD)"; the Gastroenterology paper reviews the role of imaging for early detection of PDAC; the Lancet G&H paper discusses AI's application to imaging in relation to early detection of PDAC.

Our current study focuses on model development and large-scale retrospective validation. We agree that whether the findings of this study can be translated to improved outcomes will require further prospective studies for validation (we have added this in the Discussion section [lines 661] in the revised manuscript). We believe this should be highly promising – Besides the two PDAC cases and one cured PNET reported in our original manuscript, after PANDA's deployment in our hospital SIPD, another PANDA screen-detected patient was latterly diagnosed as stage I PDAC (pT2 N0) by surgical pathology (this happened during the manuscript revision period), as detailed in the figure to answer Reviewer #3's first question.

As mentioned in a recent review (*Pickhardt, Radiology 303, 241-254, (2022)*), opportunistic screening in radiology "refers to the practice of leveraging incidental imaging data unrelated to the clinical indication, generally for the purpose of wellness, prevention, risk profiling, or presymptomatic detection of relevant disease". In our real-world study, PANDA successfully detected PDAC and nonPDAC lesions on incidental CT imaging scans unrelated to the pancreas indication (Extended Table A3), e.g., 2 PDAC (including a stage I PDAC) and 1 PNET cases on chest noncontrast CT, which were not previously detected by standard-of-care. Our real-world experiment indicated that opportunistic screening with PANDA has the potential to advance early detection of (peri-)pancreatic malignancies and high-risk lesions in the hospital population of asymptomatic patients. Opportunistic screening is considered to be more attractive than current AI triage tools because of "earlier detection of treatable disease and risk stratification for preventable disease…" (*Pickhardt, Radiology 303, 241-254, (2022)*). Nevertheless, we also understand that the definition of the term "opportunistic screening", especially for cancer, may not yet

be widely adopted. Therefore, we have added the citation of the Radiology paper and changed the heading of "opportunistic screening" in section 6 to "Feasibility of lesion detection on chest CT" following your suggestion in your question H.3.

H.2. The authors include multiple editorial comments and areas of hyperbole throughout the manuscript that should be removed. Examples are present in the PANDA methods regarding the impact of specificity (e.g., line 206), in the methods line 279-280 regarding the performance of residents in the study, line 281-5 regarding how radiologists read exams, service line considerations in chest CT (360-362), line 397 "provide striking evidence of the clinical utility…", and lines 491-492 "…the evoluted (sic) AI model is better aligned to bedside clinical needs". The manuscript should be edited to remove hyperbolic or speculative language throughout the scientific presentation. The discussion section allows space for these insights.

H.2.1. Explanatory discussion for the results should be saved for the discussion section. For example, line 374-382 hypothesizes about the reason for the performance drop in the chest CT cohort. This conjecture should be moved to the discussion.

H.2.2. Figure 4e – Speculative language that ductal dilatation alone is diagnostic of PDAC (page 12).

Response: Thank you for your valuable suggestions (H.2. – H.2.2.). We removed such editorial comments throughout the manuscript and selectively moved them into the Discussion section.

H.3. Use of term "opportunistic screening" in section 6 implies that these chest CTs were unrelated to the pancreatic abnormality, which may be confusing to readers as they were acquired at a pancreas disease center. It would be better to use a heading like "Feasibility of lesion detection on chest CT".

Response: We have changed the heading accordingly. As our response to your question C.2.2, SIPD is a center affiliated with a major tertiary hospital, so we can access the CT scans performed for various indications in the hospital. Most of these chest CTs were acquired during COVID-19 pandemic for prevention purposes.

H.4. Discussion, line 511 – "Human eyes are not sensitive to subtle imaging grayscale intensity changes". This paragraph requires grounding in the scientific literature. It current engages in misunderstanding of why intravenous contrast is used in medical imaging – it is used to improve contrast-to-noise ratios for pathology by exploiting alterations in blood flow and vessel permeability related to pathology – and cites preliminary studies on AI synthesis of contrast studies as proof that intravenous contrast is superfluous. The field of perfusion analysis research provides sufficient examples of the independent data provided by IV contrast kinetics to counteract this

speculation. The discussion of whether AI-generated images can accurately represent patient-specific pathophysiology is well beyond the scope of this paper. It would be better for the authors to focus on what their models can do with noncontrast images.

Response: We apologize for the ambiguity. This paragraph aimed to explain and provide some evidence (the cited preliminary studies on AI synthesis) why PANDA could break the performance upper-bound of human expert radiologists when read only in noncontrast CT. We have rewritten this paragraph (Section Discussion, paragraph 3) for better clarity, according to your comments. The revised paragraph removed the discussion on the indication of the related preliminary studies, and only focused on the methodological difference.

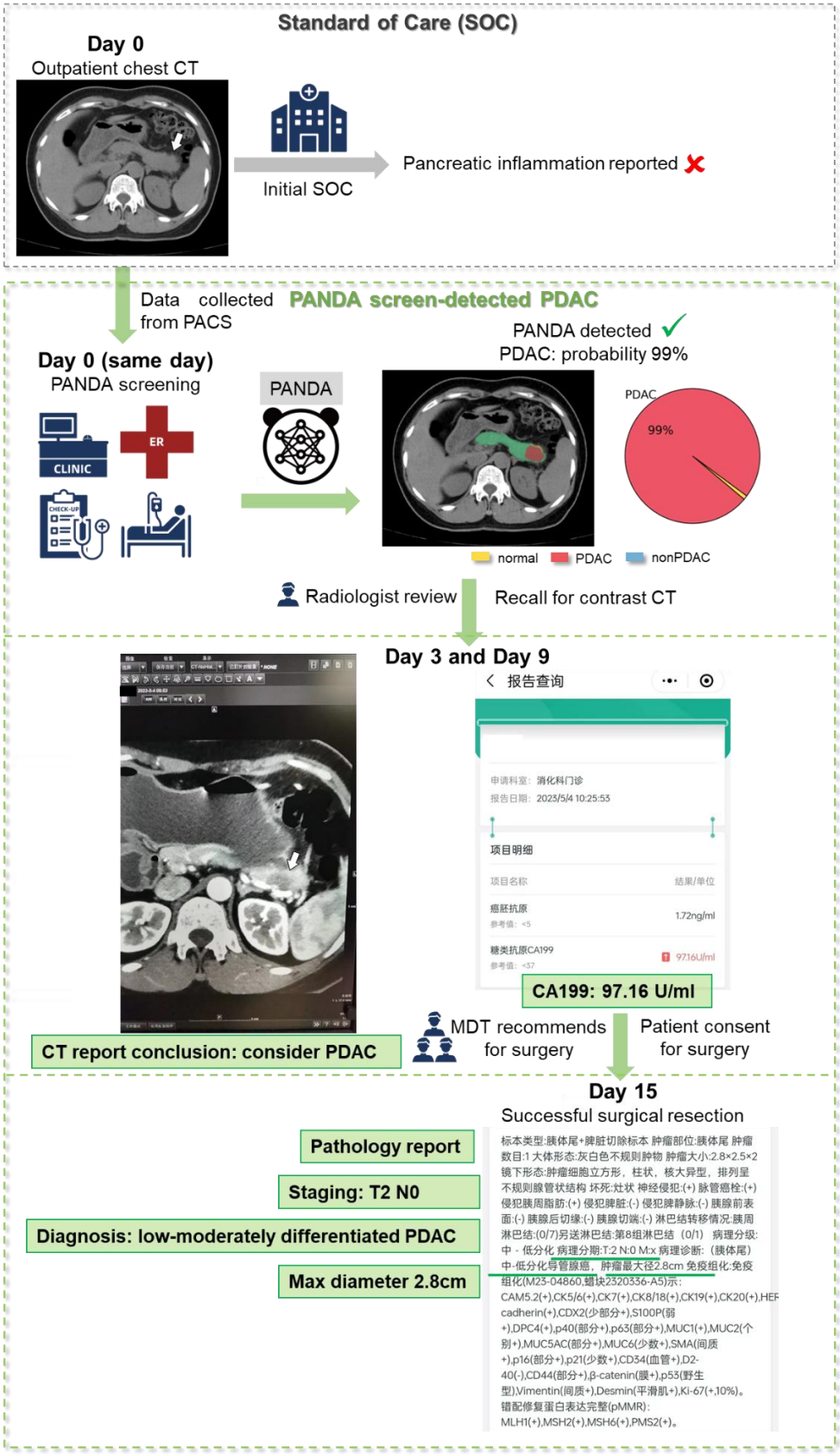
Comments from the Reviewer #3

This study presents an AI system for detection and diagnosis of pancreatic cancer based on non-contrast CT images. Overall, this is a well written manuscript with many results presented and comprehensive evaluation of model performance. Strengths of the study include addressing a clinically significant problem, developing an automated AI model for lesion detection/diagnosis, comparisons with multiple readers, multi-center validation with large patient cohorts, and real-world evaluation. However, there are also some major weaknesses around the study design, methodology, and interpretation of results as noted below.

The goal of the study seems to be overambitious. Although the AI-enabled analysis of non-contrast CT may be useful as a tool for initial detection of suspicious lesions, the diagnostic workup will always require more definitive approaches such as contrast-enhanced CT, MRI, and endoscopic US. Thinking long term about translation, relying solely on non-contrast CT while ignoring other imaging modalities for lesion diagnosis is a missed opportunity.

Response: Yes, a diagnostic workup is always required for patient management and treatment. PANDA is meant to (opportunistic or designed) screening, a pre-step before diagnosis, and therefore not to replace existing diagnostic imaging modalities.

This is also reflected by our real-world study that PANDA screen-detected lesions with potential malignancies or undetermined types were invited to undergo a contrast-enhanced MRI for diagnosis, and as a result, one such patient (nonPDAC probability 95%) was diagnosed as PNET by MRI and then by pathology after surgery (Fig. A7 in original manuscript). Similarly, after PANDA's deployment in our hospital SIPD, another PANDA screen-detected patient (PDAC probability 99%) was latterly diagnosed as PDAC by MRI and surgical pathology (this happened during the manuscript revision period), as shown in the following figure.


Standard of Care (SOC)
Day 0
Outpatient chest CT
Initial SOC
Pancreatic inflammation reported
Data collected from PACS
PANDA screen-detected PDAC
Day 0 (same day)
PANDA screening
CLINIC
ER
CHECK-UP
PANDA
PANDA detected
PDAC: probability 99%
PDAC
99%
normal
PDAC
nonPDAC
Radiologist review
Recall for contrast CT
Day 3 and Day 9
报告查询
申请科室：消化科门诊
报告日期：2023/5/4 10:25:53
项目明细
项目名称
结果/单位
癌胚抗原
参考值：<5
1.72ng/ml
糖类抗原CA199
参考值：<37
97.16U/ml
CA199: 97.16 U/ml
CT report conclusion: consider PDAC
MDT recommends for surgery
Patient consent for surgery
Day 15
Successful surgical resection
Pathology report
Staging: T2 N0
Diagnosis: low-moderately differentiated PDAC
Max diameter 2.8cm
标本类型:胰体尾+脾脏切除标本 肿瘤部位:胰体尾 肿瘤数目:1 大体形态:灰白色不规则肿物 肿瘤大小:2.8×2.5×2
镜下形态:肿瘤细胞立方形，柱状，核大异型，排列呈不规则腺管状结构 坏死:灶状 神经侵犯:(+) 脉管癌栓:(+)
侵犯胰周脂肪:(+) 侵犯脾脏:(-) 侵犯脾静脉:(-) 胰腺前表面:(-) 胰腺后切缘:(-) 胰腺切端:(-) 淋巴结转移情况:胰周淋巴结:(0/7)另送淋巴结:第8组淋巴结（0/1） 病理分级:中－低分化 病理分期:T:2 N:0 M:x 病理诊断:（胰体尾）中-低分化导管腺癌，肿瘤最大径2.8cm 免疫组化:免疫组化(M23-04860,蜡块2320336-A5)示:
CAM5.2(+),CK5/6(+),CK7(+),CK8/18(+),CK19(+),CK20(+),HEF
cadherin(+),CDX2(少部分+),S100P(弱
+),DPC4(+),p40(部分+),p63(部分+),MUC1(+),MUC2(个
别+),MUC5AC(部分+),MUC6(少数+),SMA(间质
+),p16(部分+),p21(少数+),CD34(血管+),D2-
40(-),CD44(部分+),β-catenin(膜+),p53(野生
型),Vimentin(间质+),Desmin(平滑肌+),Ki-67(+,10%)。
错配修复蛋白表达完整(pMMR):
MLH1(+),MSH2(+),MSH6(+),PMS2(+)。

As another important finding of our work, PANDA demonstrated a good initial diagnosis performance on noncontrast CT, which could provide further assistance in opportunistic screening practice. In reality, many non-pancreatic specialists face challenges in assessing the malignancy of pancreatic lesions on noncontrast CT – even when suspicious lesions are detected, they are often confused and lack enough confidence to conclude. With the initial diagnostic ability of PANDA, it can better assist physicians in triaging and managing patients based on priority levels (e.g., high, medium, low risk). As mentioned in the second example above, while the radiologist's report initially diagnosed inflammation based on noncontrast CT, PANDA provided a 99% probability of PDAC, alerting the physician and patient to give utmost attention and take fast action, resulting in surgery being completed in 15 days for a stage IB cancer. On the other hand, considering the relatively low incidence of pancreatic cancer in the general population, PANDA's initial diagnostic capability for nonPDAC may also assist physicians in reducing the risk of overdiagnosis.

We have added key points of the above discussions in the revised manuscript (Section 4 lines 651-656).

Q1: The AI model was developed using a historical case-control series. In this setting, the normal controls are severely under-sampled or underrepresented from their actual distributions (also, normal controls were randomly selected in this work; not matched with cases). The AI model would not have been exposed to a sufficiently diverse set of subjects that have 'normal' scans (see comments below). Therefore, the training set used in this study does not reflect the general population that would be screened for pancreatic cancer.

Same concerns for PDAC detection, where the training set is dominated by PDAC, while in realit most pancreatic lesions are not PDAC.

How do you make sure that AI models trained on a biased dataset generalize to the general population?

Q2: The AI model was tuned to achieve a 99% specificity for lesion detection in the case-control training cohort, in which the positive cases with lesions accounted for >70% of the cohort. On the other hand, the positive findings only represent about 1% of the population in the real-world evaluation. Reasonably, one would expect to see a drop in performance when applied to a real-world population given the 70-fold decrease in prevalence between the two scenarios. However, this is not the case with their model which maintained the same 99% specificity for lesion detection in the real-world dataset. Also, the model had a remarkably similar specificity of 98.8% in the external test set. The question is: what makes the model generalize so well to such a heterogeneous and fundamentally different population?

Response: Thank you for your comments. We understand and appreciate your concern about generalization, and we answer Q1 and Q2 together and summarize our model's generalization ability below.

(a) Diverse training population. The internal hospital (SIPD center) is a tertiary hospital with high-volume pancreatic cancer treatment in Shanghai, China. Both patients with pancreatic lesions and normal pancreas come from all parts of China, so the enrolled population of our training data is actually a diverse representation of the Chinese population. In addition, normal cases are randomly selected, which might better match our target population – patients with diverse conditions (no pancreatic lesions) undergoing CT examinations in the hospital.

(b) Noncontrast CT is more generalizable than contrast CT. CT itself is a relatively generalizable protocol for AI models, because the Hounsfield Unit (HU) values have physical meanings. A previous study, the FELIX project [31] which was recognized by Reviewer #2, have shown that even trained on only one center (Johns Hopkins Hospital) with 560 PDAC patients, 531 normal patients and 505 other pancreatic lesions, the contrast-enhanced CT-based AI can generalize well to different external centers on pancreatic lesion detection (94-97% sensitivity and 95%-99% specificity internally; 91%-97% sensitivity and 91%-93% specificity externally). Our study shows that the generalization using noncontrast CT can be even better, which could be attributed to the fact that the noncontrast CTs are not influenced by different protocols of contrast injection and phase acquisition across centers.

(c) Joint segmentation and classification model design improves generalization. Different from pure classification-based AI models (which may easily overfit to some imaging bias/shortcut), our method is a joint segmentation and classification model, where each positive classification has its local pathological basis captured by the segmentation network (trained on 2,270 lesions with 3D masks). Besides, our model adopts data augmentations to boost the generalization, i.e., random noise injection, random image contrast augmentation, random shape elastic augmentation, and the novel chest CT augmentation described in the method section. In addition, the utilization of the segmentation model allows the learning process to focus more on the limited region of the pancreas (therefore, effectively disregarding complex background variations). This, to some extent, reduces the demand for a large number of normal pancreas for training – as shown in Fig. A3a in the original manuscript, even with 10% of training data (about 90 normal pancreas), the model reaches a specificity of 98.1% on external cohorts. Moreover, in patients with pancreatic lesions, the majority of the pancreatic regions, apart from the tumor and duct, can be considered normal. This, to some extent, increases the number of normal samples available for learning – in fact, by solely utilizing the segmentation model to learn from patients with lesions, the specificity of lesion detection can reach ~80% in normal cases.

(d) Threshold adjustment to produce a 99% specificity. Our model is trained on a relatively balanced dataset (positive:negative = 7:3; 2,207 tumor and 938 normal cases). As mentioned in Method 9.3.2, we manually adjusted the probability threshold to achieve a specificity of 99% on the cross-validation of the training set and kept it fixed while testing. This threshold cut-off selection is important because a high specificity is crucial for screening of a relatively low-prevalence cancer. The multi-center validation has proved its good generalization (98.8% specificity on 1,668 normal cases). If a model can

maintain a high specificity (e.g., 99%) on multi-centers (8 centers with normal cases), this performance should be independent from the ratio of positive and negative cases, no matter tested on a balanced cohort (multi-center cohorts), or an imbalanced cohort (the real-world cohort with 1% positive). Experimentally, we showed that this technique of threshold adjustment also worked well on the real-world evaluation with mostly normal patients.

(e) Model evolution increases the specificity to 99.9%. Even with the above generalization abilities, we agree that the AI model will still experience unseen conditions in the large-scale general population that will trigger false positive predictions. That was one of the reasons why we evaluated our model in a real-world (RW) setting. PANDA maintained a specificity of 99% in the RW1 cohort. Then, the exemplary unseen conditions, e.g., pancreatic fatty infiltration and gastric content, along with false positives and false negatives from the internal, external, and RW1 cohorts, are collected and added for the model upgrade, resulting in the PANDA Plus model, which was later proved to boost the specificity to 99.5% (adjusted specificity = 99.9%) in RW2 cohort.

(f) More PDAC than nonPDAC. In the training dataset, we consecutively collected the pathology-confirmed full taxonomy of pancreatic lesions, which represent the lesion distribution under the gold standard. PDAC accounts for about 63% of these lesions. This enables a high detection (e.g., detection rate = 96.5% in external cohorts) and identification (sensitivity = 90.1% and specificity = 95.7% in external cohorts) performance for this unique lesion type – the key screening target with the most dismal prognosis. However, such data distribution results in a relatively lower performance on nonPDAC. Therefore, we did an analytical experiment by balancing the training data. Specifically, we balanced the number of each type of lesion to train the PANDA Stage-3 for differential diagnosis, where the frequency of the model seeing PDAC dropped to 12.5%. However, compared to the unbalanced version (reported in the original manuscript), the balanced accuracy dropped from 65% to 59% in the internal test cohort. Nevertheless, the performance of nonPDAC lesions still has room for improvement (e.g., adding more training data; Extended Fig. A3d).

We have condensed these discussions into the revised manuscript (Discussion section, paragraph 2, lines 539-561). Please note that we provided an interactive demo (line 91-96 in original manuscript; still active) for optional user testing. Reviewers are welcome to upload noncontrast CT data to test either the sensitivity or specificity or both by PANDA. Besides the results reported in our current manuscript, we recently have performed another retrospective real-world study in Site B (FAHZU) including ~70,000 consecutive patients' abdominal noncontrast CTs and another prospective real-world study in SIPD including ~10,000 consecutive patients' multi-scenario noncontrast CTs, PANDA Plus achieved a stable specificity of ~99.9% on both scenarios.

Q3: In the external test cohorts, the model had a sensitivity of 93.3% and 96.5% for detecting lesions and PDAC (Line 331-333). This means that 245 lesions including 95 PDAC were not

detected and missed by the AI model. However, all of the 3669 lesions and 2737 PDAC had a diagnosis prediction label in the confusion matrix for the differential diagnosis (Fig. 2f). I find these results rather peculiar. How is it possible that the model still generates a diagnosis prediction when it fails to detect there is actually a lesion in the scan?

This suggests that either there's an internal consistency in the model, or the true lesions are fed into the model after the fact, which is not how it is supposed to be evaluated or used.

Response: We apologize for the confusion. For enhanced clarity, please refer to lines 213-214, lines 995-1001, and Fig. A1a ("Stage-3") about differential diagnosis in the original manuscript. In our experiments, lesion detection and differential diagnosis were evaluated separately, not combined. Technically, PANDA Stage-3 does not require a "true lesion" for testing because this stage only takes the pancreas ROI as input, obtained from Stage-1 output. For differential diagnosis, we only evaluated PANDA Stage-3, without considering whether or not PANDA Stage-2 detected a certain lesion. One reason is that the differential diagnosis module could be independently evaluated following the pancreatic tumor/cyst classification task (*Springer et al, Science Translational Medicine 11, eaav4772, (2019)*; *Chu et al, Abdominal Radiology 47, 4139-4150, (2022)*) without normal patients included and each patient will have a lesion type assigned. Another reason is that in such a way, we could compare our diagnosis results to the radiology report, in which a bias existed that our preoperative radiology report will not miss any lesions, i.e., 100% sensitivity in lesion detection (because our inclusion criteria is patients undergoing surgical resection, their preoperative CTs are for surgical planning after clinical diagnosis), though we acknowledge that this may not be completely true in reality.

Inspired by your suggestion, we now provide the confusion matrix of the full pipeline, i.e., 9-way classification including normal, PDAC, and 7 types of nonPDAC lesions, where we achieved an accuracy of 79.8% (95% CI 76.9-82.5%) and a balanced accuracy of 62.1% (95% CI 57.7-66.3%) internally on the combination of internal test cohort and internal additional cohort (a newly added cohort in an effort to address your question Q4); and an accuracy of 84.7% (95 CI 83.7-85.7%) and a balanced accuracy of 52.2% (95% CI 50.0-54.4%) externally.

We have added these further explanations in the revised manuscript (Section 5.4.3) and added the full pipeline's results (lines 255-257; lines 370-371; Supplementary Fig. B13).

Q4: The internal evaluation set is too small to evaluate differential diagnosis, where the number of lesions comes down to single digits for several lesion types.

Response: Thanks for pointing this out. We newly added 611 patients treated between November 2020 to October 2021 at the SIPD. These patients, along with the original internal test cohort patients, constitute a new cohort named the internal differential diagnosis cohort. Now, the minimal number of lesions (i.e., MCN) is 23. Please refer to Section 2.2 for the results of the new cohort.

Q5: The authors claim in L527: 'PANDA is an interpretable deep model'. However, no interpretation results were shown to understand exactly what types of image features drive the prediction for lesion classification. In other words, what imaging characteristics specifically does the AI model use that can help detect and distinguish different lesion types on non-contrast CT scans that are difficult to discern even for expert radiologists?

Response: Thank you for the valuable suggestion. For the analysis of interpretability, we visualized the Grad-CAM heatmap of PANDA Stage-2, to see which part of the image contributes most to the classification of abnormality. We also visualized the top activated attention maps of the Transformer branch of PANDA Stage-3 to interpret how PANDA classified the lesions. The memory tokens of the Transformer not only attended the lesion locations but also considered the secondary signs for lesion differential diagnosis as utilized by the radiologists. This analysis has been added to the revised manuscript in Section 2.2 (lines 278-280) and Section 5.7 (lines 1295-1306). The highlight of the analysis was shown in Extended Data Fig. A4.

Q6: Some descriptions of the key methodology seem counterintuitive, are rather vague and lack sufficient details. For instance, in Line 201-3: it is stated that 'Memory Transformer branch that was used to automatically encode the feature prototypes of the pancreatic lesions, such as local textures, positions, and pancreas shapes.' How do positions inform diagnosis given that PDAC and other lesions can appear anywhere in the pancreas?

Line 873-4: 'The memory branch starts with learnable memories designed to store both positional and texture-related prototypes of the eight types of pancreatic lesions.' What are these positional and texture-related lesion prototypes and how are they defined exactly?

Response: First, different types of pancreatic lesions have their specific position distribution patterns within the pancreas. For example, as reported in an international multicenter study (*Springer et al., Science Translational Medicine 11, eaav4772, (2019)*) and our previous study of SIPD patients (*Zhao et al., In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 13743–13752 (2016); https://arxiv.org/abs/2012.04701*), nearly all (e.g., 97%) MCN lesions occur in the body/tail of the pancreas; the majority of (e.g., 61%-79%) SPT, PNET, and SCN lesions are also found in the body/tail of the pancreas; on the other hand, IPMN lesions are more commonly (e.g., 57%) located in the head/uncinate of the pancreas. This positional information can aid the diagnosis of pancreatic lesions. Second, the positional relationship between lesion and duct is another cue for diagnosis. For example, PDAC usually causes and connects to a dilated pancreatic duct; IPMN originates inside the pancreatic duct, so the location of the lesion to the pancreatic duct is useful to differentiate IPMN and its subtype (main- or branch-duct). Third, as shown in our newly added model interpretability analysis (your above question Q5; Extended Data Fig. A4), positional information combined with textural information is the primary cue utilized by PANDA and radiologists for diagnosis.

Regarding the positional and texture-related lesion prototypes, we clarify these components here and in the revised manuscript (lines 1047-1053). These prototypes are implemented as learnable

memory vectors $x^m \in R^{N\times d}$ with a length of N and a dimension of d in the memory branch (Extended Fig. A1) and are automatically updated during the training process. Positional information is encoded into the learnable memory vectors via positional embeddings and textual features are extracted by the CNN branch. Specifically in our architecture, given a 3D feature map $x^c \in R^{H_0\times W_0\times D_0\times d_0}$ of the UNet branch with the shape of (H_0, W_0, D_0) and d_0 channels, we first tri-linearly interpolate x^c into a fixed shape (H, W, D) and linearly project it to d dimension. We then flatten it into 1D features of a length of $N = H \times W \times D$ and add a learnable positional embedding $x_{pos} \in R^{N\times d}$ shared among each layer. The resulted 1D feature $\hat{x}^c \in R^{N\times d}$ thus encodes both the texture-related feature from the UNet branch and the position-related feature from the positional embeddings. The positional embeddings were implemented as randomly initialized learnable parameters, each associated with a specific position in the image. Through the self-attention and cross-attention mechanism described in Method 9.3.3 Eq. (2), both the positional information and the textual information are encoded into the learnable memory vectors, and automatically updated during training.

The idea of learnable memory vectors (a.k.a. object queries) that leverages a CNN feature extractor and positional embeddings were developed in recent success AI architectures in computer vision, such as DETR (*Carion et al., 2020 European Conference on Computer Vision, 213-229 (2020)*), MAX-Deeplab (*Wang et al. , 2021 IEEE/CVF conference on computer vision and pattern recognition, 5463-5474 (2021)*), and MaskFormers *(Cheng et al., 2021 Advances in Neural Information Processing Systems, 17864-17875 (2021*)) for the purpose of visual instance detection, segmentation and classification, which are closely related to our task of pancreatic lesion detection, segmentation, and differential diagnosis. To interpret these memory vectors in our model, we visualized their attention maps which showed the position that these memory vectors attended to. As shown in Extended Data Fig. A4, some memory vectors were attended to tumor locations as response for local texture changes, some on pancreas characteristics, e.g., pancreatic atrophy and dilated duct, and others on position-related cues, e.g., relationship to pancreatic duct, which jointly assisted the model to make a final diagnosis.

Minor:

Q7: What are the distributions of CT acquisition parameters such as slice thickness, mAs levels, etc? Do they affect lesion detection or diagnosis by the AI model?

Response: More information on CT acquisition parameters has been added in the revised manuscript (Extended Table 1) following your and Reviewer #2's comments. Our model shows robust lesion detection performance (Fig. 2a-d) across multiple centers with different CT acquisition parameters, e.g., AUC = 0.996 internally and 0.984 externally. Regarding lesion diagnosis, since individual cohort/center may have a small number of certain lesions, it is more reasonable to compare between the two large cohorts, i.e., internal training set (evaluated by cross-validation; Fig. A2b; n=3,208) vs. all external test cohorts (Fig. 2f; n=3,669). The balanced accuracy has only a slight drop of 1.6% (54.2% vs. 52.6%). We have revised the manuscript accordingly (line 374).

Q8: Segmentation of the pancreas on non-contrast is a challenging problem by itself. What is the model performance for segmenting pancreas and different types of lesions?

Response: Thank you for the question. Please refer to lines 274-278 and Supplementary Table B8 in the revised manuscript for the segmentation accuracy.

Q9: The study includes mostly patients from a Chinese population and a small proportion of patients from one Eastern European country. It's an overstatement to claim this is an 'international' validation study on multiple occasions.

Response: Thank you for highlighting this. This is a shared question with Reviewer #1 ("only 4% of the multi-center cohort comes out of China/Taiwan"). We changed the name of the cohort from "international multicenter cohort" into "external multicenter cohort" throughout the paper and added this limitation to Discussion section (lines 657-661).