frontiers | Frontiers in Radiology

TYPE Original Research
PUBLISHED 11 July 2023
DOI 10.3389/fradi.2023.1202412

Check for updates

OPEN ACCESS

EDITED BY
Samuel Kadoury,
Polytechnique Montréal, Canada

REVIEWED BY
Lorenzo Faggioni,
University of Pisa, Italy
Bulat Ibragimov,
University of Copenhagen, Denmark

*CORRESPONDENCE
David Dreizin
✉ daviddreizin@gmail.com

RECEIVED 08 April 2023
ACCEPTED 22 June 2023
PUBLISHED 11 July 2023

CITATION
Dreizin D, Zhang L, Sarkar N, Bodanapally UK, Li G, Hu J, Chen H, Khedr M, Khetan U, Campbell P and Unberath M (2023) Accelerating voxelwise annotation of cross-sectional imaging through AI collaborative labeling with quality assurance and bias mitigation.
Front. Radiol. 3:1202412.
doi: 10.3389/fradi.2023.1202412

COPYRIGHT
© 2023 Dreizin, Zhang, Sarkar, Bodanapally, Li, Hu, Chen, Khedr, Khetan, Campbell and Unberath. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Accelerating voxelwise annotation of cross-sectional imaging through AI collaborative labeling with quality assurance and bias mitigation

David Dreizin[1]*, Lei Zhang[1], Nathan Sarkar[1], Uttam K. Bodanapally[1], Guang Li[1], Jiazhen Hu[2], Haomin Chen[2], Mustafa Khedr[1], Udit Khetan[1], Peter Campbell[1] and Mathias Unberath[2]

[1]Department of Diagnostic Radiology and Nuclear Medicine, School of Medicine, University of Maryland, Baltimore, MD, United States, [2]Johns Hopkins University, Baltimore, MD, United States

**Background:** precision-medicine quantitative tools for cross-sectional imaging require painstaking labeling of targets that vary considerably in volume, prohibiting scaling of data annotation efforts and supervised training to large datasets for robust and generalizable clinical performance. A straight-forward time-saving strategy involves manual editing of AI-generated labels, which we call AI-collaborative labeling (AICL). Factors affecting the efficacy and utility of such an approach are unknown. Reduction in time effort is not well documented. Further, edited AI labels may be prone to automation bias.

**Purpose:** In this pilot, using a cohort of CTs with intracavitary hemorrhage, we evaluate both time savings and AICL label quality and propose criteria that must be met for using AICL annotations as a high-throughput, high-quality ground truth.

**Methods:** 57 CT scans of patients with traumatic intracavitary hemorrhage were included. No participant recruited for this study had previously interpreted the scans. nnU-net models trained on small existing datasets for each feature (hemothorax/hemoperitoneum/pelvic hematoma; $n = 77–253$) were used in inference. Two common scenarios served as baseline comparison- *de novo* expert manual labeling, and expert edits of trained staff labels. Parameters included time effort and image quality graded by a blinded independent expert using a 9-point scale. The observer also attempted to discriminate AICL and expert labels in a random subset ($n = 18$). Data were compared with ANOVA and post-hoc paired signed rank tests with Bonferroni correction.

**Results:** AICL reduced time effort 2.8-fold compared to staff label editing, and 8.7-fold compared to expert labeling (corrected $p < 0.0006$). Mean Likert grades for AICL (8.4, SD:0.6) were significantly higher than for expert labels (7.8, SD:0.9) and edited staff labels (7.7, SD:0.8) (corrected $p < 0.0006$). The independent observer failed to correctly discriminate AI and human labels.

**Conclusion:** For our use case and annotators, AICL facilitates rapid large-scale curation of high-quality ground truth. The proposed quality control regime can be employed by other investigators prior to embarking on AICL for segmentation tasks in large datasets.

KEYWORDS

AI assisted annotation, AI-collaborative labeling, artificial intelligence—AI, CT volumetry, computed tomography, trauma, quantitative visualization, human in the loop (HITL)

## Introduction

Scalable artificial intelligence solutions require large datasets with high quality annotation, and for quantitative visualization tools, this has traditionally entailed prohibitively painstaking manual voxelwise labeling (1, 2). To date, few FDA-approved commercially available tools perform segmentation and quantification tasks (3–5). In 2018, The NIH, RSNA, ACR, and the Academy for Radiology and Biomedical Imaging Research held a workshop that stressed data scarcity as a major obstacle requiring approaches for high-throughput data curation (6) and spurred research efforts aimed at accelerating annotation in the domains of federated learning, active learning, artificial intelligence-assisted annotation (AIAA) with limited clicks and scribbles, semi-supervised learning, and synthetic data augmentation (7–18).

A simple AI collaborative labeling (AICL) method distinct from these approaches, involves training on a subset of manually labeled seed data, and then editing inference labels on unseen studies. Expert editing of automated labels was performed as a final annotation step for the public COVID 19–20 Lung CT Lesion Segmentation Challenge (19) dataset. Substantial time savings may be possible, especially for studies with extensive infiltrates, as studies with larger target volumes take longer to label (13). Another application of this collaborative approach involves editing inference labels and iteratively retraining a model in batches (20, 21). Time savings result with each batch as the model becomes more generalizable and there are fewer errors to correct. With these approaches, automation and complacency bias become potential concerns (22)- for example, labelers may be lulled into assuming automated labels are correct or second guess their own correct judgement.

A broader issue is the general fallibility of expert labelers (23) and the near-universal assumption in the published literature that manual ground truth is the ideal benchmark for determining algorithm performance (24). Taking COVID infiltrates as an example, annotators are faced with multifocality; highly variable shapes, locations, and volumes of infiltrates; widely differing attenuation values; irregular contours; and adjacent opacities such as atelectasis that can be mistaken for infiltrate. Even when using inference labels as a starting point, a mean Dice Similarity Coefficient (DSC) of only 0.70 was achieved between observers in the aforementioned COVID challenge (19). In a study of traumatic hemoperitoneum, a similarly complex task, mean test-retest DSC for manual annotator was also 0.70 (25). Algorithms will tend to ignore outliers in new data, and this is typically acknowledged as a flaw with respect to robustness. But when considering outliers related to mislabeling by humans, this vice of overfitting can become a virtue, as human inconsistencies may be ignored to arrive at a more uniformly satisfactory label (**Figure 1**). Thus, there is a possibility that an edited inference label with less-than-perfect overlap to a manual label may be qualitatively more suitable for use as ground truth.

As a "sanity check", a qualitative or semi-quantitative label quality assessment could be performed by a blinded expert

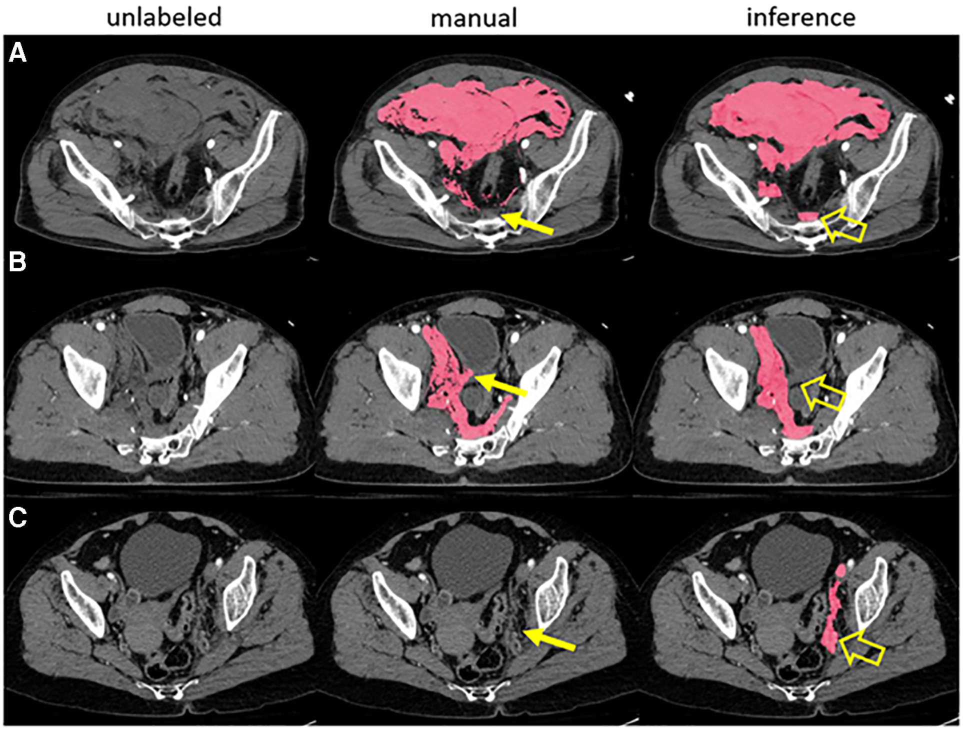

FIGURE 1
A prior pelvic hematoma segmentation algorithm correctly labeled hemorrhage in three examples with human error in manual training data (A-C). The expert missed a small volume of presacral blood in case A, inadvertently labeled a small segment of bladder wall in case B and missed a small volume of left pelvic sidewall hematoma in case C (thin arrows). The model, trained in 5-fold cross-validation, labeled these areas correctly in inference (open arrows).

observer to determine whether manual or automated labels are the better ground truth for a given study. Semi-quantitative scores that are higher for edited automated labels empirically exclude detrimental automation bias. We are not aware of prior work in medical imaging segmentation that has employed this approach.

A recent scoping review of AI CAD tools for trauma found that while very large datasets have been developed for 2D detection tasks such as for pulmonary abnormalities and extremity factures on plain radiography, dataset sizes for torso hemorrhage-related pathology ranged between fewer than 100 to 778 patients (4, 25–31). Despite the perceived unmet need for hemorrhage-related quantitative visualization tools (32), due to challenges unique to the torso- namely small target to total volume ratios and complex pathology- deep learning (DL) algorithms yielding meaningful visual results for the chest, abdomen, and pelvis, have been late-comers in medical image analysis (33). A limited number of studies using small datasets have demonstrated an association between intracavitary torso hemorrhage volumes and clinical outcomes (25, 29, 34–36), with comparable or improved prediction compared to existing categorical grading systems. In 2021, Isensee et al., introduced nnU-net (37), an out-of-the-box self-configuring data-driven method which incorporates a 3D coarse-to-fine multiscale approach that has shown state-of-the-art performance on numerous public medical imaging datasets over bespoke multiscale methods.

A rapid-throughput labeling strategy using an AICL approach could make high-quality voxelwise annotation feasible for very large corpuses of CT data. Using 57 unseen trauma torso CT exams, we conducted this pilot study to determine time savings and annotation quality associated with AICL using nnU-net models trained with existing labeled seed data for several hemorrhage-related features- namely hemothorax, hemoperitoneum, and extraperitoneal pelvic hematoma. Two commonly employed labeling approaches- manual expert labeling, and manual labeling by trained staff with expert editing- are used for comparison. Label quality is compared using blinded assessment with a proposed 9-point RAND/UCLA Likert grading system.

# Methods

## Deep learning method

We trained nnU-net (37) in five-fold cross-validation using three existing datasets with voxelwise labeling of the feature of interest from previous pilot studies- a hemoperitoneum dataset of 130 patients, a pelvic hematoma dataset of 253 patients, and a hemothorax dataset of 77 patients. Cascaded (3D low-resolution to high-resolution) nnU-net yielded improved DSCs over prior bespoke methods for all three use cases (25, 29, 34, 38), with mean DSC increasing to 0.75 from 0.61 for hemothorax; to 0.67 from 0.61 for hemoperitoneum, and to 0.75 from 0.71 for pelvic hematoma.

## CT imaging

Trauma whole-body CT studies are performed at our institution on one of two trauma bay-adjacent scanners- a 64-section scanner (Brilliance; Philips Healthcare, Andover Mass- scanner 1), and a dual-source 128-section scanner (Siemens Force; Siemens, Erlangen, Germany- scanner 2). Studies are performed with 100 ml of intravenous contrast material [Omnipaque (iohexol; 350 mg/ml iodine)] at an injection rate of 5 ml/sec (scanner 1) or 6 ml/sec (scanner 2) for the first 60 ml followed by 4 ml/sec for the remainder. For penetrating trauma, we routinely administer rectal contrast (50 ml iohexol with 300 mg of iodine diluted in 1l of water) if abdominal penetration is suspected based on the location of entry and/or exit wounds. Arterial phase images are acquired from the thoracic inlet through the greater trochanters, followed by portal venous phase images through the abdomen and pelvis beginning just above the dome of the diaphragm. Images are archived at 3 mm section thickness.

## Dataset

Our primary aim was to determine time savings using AICL on new patients. For time analyses using comparison of means, we expected large effect sizes. Cohen recommended a D value of 0.2, 0.5, and 0.8 for low, medium, and high effect sizes respectively (39). A pre-hoc power calculation showed that for a power of 0.8, alpha of 0.05, and Cohen's D of 0.8, 15 patients would be needed, and with power increased to 0.9, 19 patients would be needed. Therefore, we created a convenience sample dataset of 19 randomly selected previously unseen adult (age ≥ 18) patients for each injury type from the period between June 2019 and September 2022. The total study sample included 57 patients [42 (74%) male, 15 female], with median age of 33 [IQR: 24, 58]. Forty-seven (82%) had blunt, and 10 (18%) had penetrating injury mechanisms.

## Artificial intelligence-collaborative labeling (AICL) vs. human labeling

Human labeling strategies reflected common research practices- either (a) *de novo* annotation (i.e., no AI assistance) by a radiology attending expert with over 7 years of experience with voxelwise labeling of CT datasets, or (b) *de novo* annotation by trained research staff (a volunteer third-year medical student with one year of prior cross-sectional quantitative imaging experience), with editing of staff labels by the attending. Seed data-trained nnU-net models were used to generate labels for the 57 unseen studies, and these were also subsequently edited by the expert- a process we refer to as AI-collaborative labeling. A minimum washout period of two weeks was used between sessions, and studies were labeled in random order. De novo human labeling time and attending editing times were recorded.

Labeling was performed using 3D slicer (version 5.0.3; www.slicer.org) with the spherical threshold paint tool in the range of −20–100 Hounsfield Units to minimize noise while avoiding voxels with neighboring fat, lung, or bone. At the margins of structures with similar density, the brush size is changed or the image zoomed in as needed to ensure the label conforms to the desired contour. For hemothorax, labeling and AI inference were performed in the arterial phase as images of the chest are only acquired in this phase. For hemoperitoneum and pelvic hematoma, labeling and AI inference were performed in the portal venous phase. Labeling was conducted primarily in the axial plane with supervision and refinement in the three orthogonal planes as needed. CT volumes and segmentation masks were saved in the NIfTI and NRRD file formats respectively.

## Likert grading

A radiologist domain expert with 15 years of trauma experience was shown (1) *de novo* expert labels (2) *de novo* staff labels (3) staff labels with expert edits, (4) unedited nnU-net ("AI only") labels, and (5) nnU-net labels with attending edits (AICL results). The radiologist was aware that, given the aim of the study, some results may involve AI and some would not but was blinded to the type and provenance of label sources and studies were shown in random order. The goal of voxelwise labeling is to achieve a trustworthy verifiable result that could be used at the clinical point of care. We define labels that meet these criteria as reaching a standard of "clinical quality". The blinded radiologist assigned a Likert grade based on a UCLA/RAND 9-point scale, with a score of 1–3 indicating poor label quality (this was defined as requiring substantial edits for a large number of either false positive or false negative voxels, or ignoring the label and starting again from scratch), a score of 4–6 indicating moderate label quality (requiring further editing to reach clinical quality), and a score of 7–9 indicating excellent label quality requiring minimal to no edits (Table 1). The latter range is considered a clinically acceptable result. The first six consecutive patients were randomly selected from each group (hemothorax, hemoperitoneum, and pelvic hematoma) to determine whether the radiologist could distinguish between AI-only labels, and *de novo* expert labels.

## Correlations: label times, training cases, and volumes

We sought to determine to what degree *de novo* labeling times were correlated with label volumes (in milliliters) and whether attending expert editing times for staff or automated labels were decoupled from volumes. To assess internal validity of the Likert grading scale, we compare AICL to staff labels in terms of difference in Likert scales, dice similarity coefficient, and editing time- wherein higher differences in grade were expected to correlate with lower DSCs and longer edits.

## Statistical analysis

Analysis of variance (ANOVA) was used to determine a statistically significant signal within repeated measurements. Post-hoc paired signed rank tests were used to compare segmentation time and Likert grade differences. Bonferroni correction was performed to account for multiple comparisons. Correlations between label volumes and times were assessed using Pearson *r*. Spearman correlation was used for comparison involving categorical Likert grading. *p*-values < 0.05 were considered to indicate statistical significance. Statistical analysis was performed using STATA software (version 15.1).

# Results

## Results in the total sample

Median labeling times for the total sample ($n = 57$) were 788 s (s) [IQR: 525, 1,418] for *de novo* staff labels, 602 s [IQR: 368, 1,092] for *de novo* expert labels, 193 s [IQR: 113, 269] for expert edits of staff labels, and 69 s [IQR: 48, 106] for AICL (Table 2). ANOVA yielded a *p*-value < 0.0001. The AICL approach (expert edits of automated labels) reduced time effort compared to editing of staff labels by a factor of 2.8, (*p*-value < 0.0006; note: Bonferroni corrected *p*-values) (Table 3). AICL was faster than *de novo* labeling by a factor of 8.7 (expert) to 11.4 (staff) (both *p*-values < 0.0006). Examples of AI-only labeling errors addressed in AICL are shown in Figure 2. Mean Likert grades for the total sample were 8.4 (SD: 0.6) for AICL, which was significantly higher than for *de novo* expert labels [7.8 (SD 0.9), $p < 0.0006$], for edited staff labels [7.7 (SD: 0.8), $p < 0.0006$], for AI-only labels [7.8 (SD: 1.0), $p = 0.001$], and for staff labels [6.3 (SD 1.5), $p < 0.0006$]. In the 18-patient subset (6 per intracavitary hemorrhage feature, Likert score range: 5–9) presented for discrimination of AI-only and *de novo* expert labels, 12 patients were misclassified by the independent expert observer, yielding a 66% discordance rate. Correctly and incorrectly classified examples are shown in Figures 3A–C. Using the best-case scenario, AICL, as the reference standard, and worst-case scenario (staff only labels) for comparison, the difference in Likert score was inversely correlated with the DSC with r of −0.55 ($p < 0.0001$) and directly correlated to attending editing times for

TABLE 1 Segmentation label quality Likert grading system (9-point RAND/UCLA scale).

| Score range | Description |
|---|---|
| Poor quality<br>(Score 1–3) | Mostly erroneous label that does not elicit user trust. Large number of either false positive or false negative voxels. May require discounting the label and starting again given time effort required to delete false positive portions of the segmentation. |
| Moderate quality<br>(Score 4–6) | Errors negatively impact interpretability/explainability and trust in volumetric results. Require further editing to reach clinically acceptable quality. |
| Excellent quality<br>(Score 7–9) | Clinically acceptable label quality with explainable/high trust results. An optimal segmentation result is achievable with minimal to no editing. |

TABLE 2 Summary results for total sample and individual features.

| Statistic | Vol.[a] (ml) | Time effort (s) | | | | Likert quality grade (1–9) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Expert | Staff | Staff edits | AICL | Expert | Staff | Staff edits | AI-only | AICL |
| Total sample (*n* = 57) | | | | | | | | | | |
| Mean | 362 | 752 | 1,092 | 200 | 90 | 7.8 | 6.3 | 7.7 | 7.8 | 8.4 |
| Std dev | 313 | 468 | 802 | 106 | 74 | 0.9 | 1.5 | 0.8 | 1.0 | 0.6 |
| Median | 274 | 602 | 788 | 193 | 69 | 8 | 6 | 8 | 8 | 8 |
| IQR | [108, 494] | [368, 1,092] | [525, 1,418] | [113, 269] | [48, 106] | [7, 9] | [6, 7] | [7, 8] | [7, 9] | [8, 9] |
| Min | 28 | 197 | 304 | 44 | 23 | 6 | 1 | 6 | 5 | 7 |
| Max | 1,226 | 2,139 | 4,162 | 512 | 526 | 9 | 9 | 9 | 9 | 9 |
| Hemothorax (*n* = 19) | | | | | | | | | | |
| Mean | 293 | 561 | 594 | 147 | 102 | 8.1 | 7.5 | 7.9 | 7.3 | 8.5 |
| Std dev | 315 | 378 | 256 | 105 | 46 | 0.8 | 1.0 | 0.8 | 0.8 | 0.6 |
| Median | 130 | 422 | 537 | 101 | 99 | 8 | 7 | 8 | 7 | 9 |
| IQR | [103, 384] | [315, 777] | [399, 725] | [75, 178] | [59, 135] | [7, 9] | [7, 8] | [7, 8] | [7, 8] | [8, 9] |
| Min | 60 | 197 | 304 | 44 | 42 | 7 | 5 | 6 | 5 | 7 |
| Max | 1,226 | 1,749 | 1,429 | 445 | 186 | 9 | 9 | 9 | 8 | 9 |
| Hemoperitoneum (*n* = 19) | | | | | | | | | | |
| Mean | 431 | 890 | 1,731 | 241 | 113 | 7.4 | 5.2 | 7.3 | 7.7 | 8.3 |
| Std dev | 328 | 418 | 932 | 87 | 109 | 1.0 | 1.6 | 0.6 | 1.1 | 0.5 |
| Median | 323 | 802 | 1,764 | 236 | 84 | 7 | 6 | 7 | 8 | 8 |
| IQR | [152, 723] | [549, 1,346] | [929, 2,313] | [185, 270] | [49, 128] | [7, 8] | [5, 6] | [7, 8] | [7, 9] | [8, 9] |
| Min | 39 | 277 | 537 | 126 | 29 | 6 | 1 | 6 | 6 | 7 |
| Max | 1,063 | 1,651 | 4,162 | 512 | 526 | 9 | 7 | 9 | 9 | 9 |
| Pelvic Hematoma (*n* = 19) | | | | | | | | | | |
| Mean | 364 | 805 | 951 | 213 | 54 | 8.1 | 6.2 | 7.8 | 8.5 | 8.5 |
| Std dev | 279 | 532 | 566 | 101 | 20 | 0.8 | 0.9 | 0.9 | 0.6 | 0.6 |
| Median | 312 | 602 | 830 | 221 | 53 | 8 | 6 | 8 | 9 | 9 |
| IQR | [97, 494] | [363, 1,133] | [451, 1,245] | [119, 277] | [35, 76] | [7, 9] | [6, 7] | [7, 8] | [8, 9] | [8, 9] |
| Min | 28 | 259 | 330 | 57 | 23 | 7 | 5 | 6 | 7 | 7 |
| Max | 929 | 2,139 | 2,086 | 446 | 88 | 9 | 8 | 9 | 9 | 9 |

[a] Volume is provided based on calculations from AICL.
ml, milliliters; s, seconds; Std dev, 1 standard deviation; IQR, interquartile range [Q1, Q3]; Min, minimum; Max, maximum, AICL, artificial intelligence-collaborative labeling. "staff edits" refers to expert editing of staff labels.

the staff labels with r of 0.42 ($p = 0.001$). These correlations are considered "moderate" using the Dancey and Reidy qualitative schema (40). For a Likert difference ≥4, 3, 2, 1, and 0, mean DSCs were 0.49, 0.67, 0.69, 0.65, and 0.83 respectively.

## Hemothorax

nnU-net models were trained on manually labeled seed data with 77 patients. Among the 19 unseen hemothorax studies, AI-collaborative labels required less time effort and received the highest quality Likert grades compared with other methods. Median AICL time was 99 s [IQR: 59–135]. This was comparable to median editing time for staff labels 101 s [IQR: 75–178, $p =$ 0.32], but was 4.3-fold faster than *de novo* expert labeling [422 s (IQR: 315–777), $p = 0.0006$], and 5.4-fold faster than *de novo* staff labeling [533 s (IQR: 399–725), $p = 0.0006$] (see **Tables 2, 3**). Likert scores for AI-collaborative labeling (mean: 8.5, SD: 0.6) were significantly higher than AI-only Likert grades (mean: 7.3, SD: 0.8, $p = 0.002$) and *de novo* staff labels (mean: 7.5, SD: 1.0, p = 0.004), and higher but comparable to *de novo* expert labels (8.1, SD: 0.8, $p = 0.49$). All AI-collaborative labels, *de novo* expert labels, and edited staff labels received Likert scores in the excellent range (scores of 7–9), whereas 3 automated labels and 2 staff labels received scores in the moderate range (scores of 4–6). Of the six studies presented for discrimination between automated and human labeling, the independent blinded observer believed that two AI-only labels were the product of human labeling, and that two of the expert labels were AI-only labels.

## Hemoperitoneum

nnU-net models were trained on labeled seed data in 130 patients. Among the 19 unseen hemoperitoneum studies, AI-collaborative labels again required the least time effort and received the highest quality Likert grades. AICL required median time effort of 84 s [IQR: 49–128]). Significantly more time effort (approximately 2.8-fold difference) was required for attending edits of staff labels [Median: 236 s (IQR: 185–270), $p = 0.011$]. Time effort for *de novo* expert [Median: 802 s (IQR: 549–1346), $p = 0.0006$], and *de novo* staff labeling [Median: 1,764 sec (IQR: 929–2,313), $p = 0.0006$] were greater than AICL by a factor of 9.5 and 21 respectively (see **Tables 2, 3**). Likert grades for AICL (8.3,

TABLE 3 Corrected *p*-values for multiple comparisons- total sample and subanalyses.

| Statistic | Time effort (s) | | | | Likert quality grade (1–9) | | | |
|---|---|---|---|---|---|---|---|---|
| | Staff | Staff ed. | AICL | | Staff | Staff ed. | AI-only | AICL |
| Total sample (*n* = 57) | | | | | | | | |
| Expert | 0.002 | <0.0006 | <0.0006 | Expert | <0.0006 | 1.0 | 1.0 | <0.0006 |
| Staff | | <0.0006 | <0.0006 | Staff | | <0.0006 | <0.0006 | <0.0006 |
| Staff ed. | | | <0.0006 | Staff ed. | | | 1.0 | <0.0006 |
| | | | | AI-only | | | | 0.001 |
| Hemothorax (*n* = 19) | | | | | | | | |
| Expert | 1 | 0.0006 | 0.0006 | Expert | 0.274 | 1.0 | 0.06 | 0.493 |
| Staff | | 0.0006 | 0.0006 | Staff | | 0.19 | 1.0 | 0.004 |
| Staff ed. | | | 0.3204 | Staff ed. | | | 0.064 | 0.029 |
| | | | | AI-only | | | | 0.002 |
| Hemoperitoneum (*n* = 19) | | | | | | | | |
| Expert | 0.0066 | 0.0012 | 0.0006 | Expert | 0.002 | 1.0 | 1.0 | 0.02 |
| Staff | | 0.0006 | 0.0006 | Staff | | 0.001 | 0.002 | 0.001 |
| Staff ed. | | | 0.0114 | Staff ed. | | | 1.0 | 0.001 |
| | | | | AI-only | | | | 0.249 |
| Pelvic hematoma (*n* = 19) | | | | | | | | |
| Expert | 0.8844 | 0.0012 | 0.0006 | Expert | 0.002 | 1 | 0.477 | 0.299 |
| Staff | | 0.0006 | 0.0006 | Staff | | 0.004 | 0.001 | 0.001 |
| Staff ed. | | | 0.0006 | Staff ed. | | | 0.053 | 0.018 |
| | | | | AI-only | | | | 1.0 |

AI-only refers to inference labels without editing.
Staff edits (abbreviated "Staff ed") refer to expert edits of staff labels.
All *p*-values are corrected using Bonferroni method such that e.g., a *p*-value of <0.0001, becomes <0.0006.

SD: 0.5) were significantly higher than those for attending labels (mean: 7.4, SD: 1.0, $p = 0.02$), edited staff labels (mean: 7.3, SD: 0.6, $p = 0.001$), and *de novo* staff labels (mean: 5.2, SD: 1.6, $p = 0.001$). Differences between AICL and AI-only labels (mean: 7.7, SD: 1.1) did not reach significance ($p = 0.25$). All AICL labels had excellent-range Likert scores. AI-only labels, and *de novo* attending labels received scores in the moderate range in four instances each, whereas *de novo* staff labels received moderate scores in 13 studies, and poor-quality scores (range: 1–3) in three. Of the six studies presented for discrimination between automated and human labeling, the independent blinded observer again believed that two of the AI-only labels were labeled by the expert, and that two of the expert labels were AI-only labels.

## Pelvic hematoma

nnU-net models were trained on labeled seed data in 253 patients. Among the 19 patients, AICL annotations required the least time effort [Median: 53 s (IQR: 35–76)], and were 4.2-fold, 11.4-fold, and 15.7-fold faster than edited staff labels [Median: 221 s (IQR: 119–277), $p = 0.0006$], *de novo* expert labels [Median: 602 s (IQR: 363–1,133), $p = 0.0006$], and *de novo* staff labels [Median: 830 s (IQR: 451–1,245), $p = 0.0006$], respectively (see Tables 2, 3). AI-only and AICL annotations received the same mean Likert grades (Mean: 8.5, SD: 0.6). All scores for these two groups were in the excellent range (score 7–9) with discrepancy by 1 Likert grade in 4 studies, and a discrepancy by 2 grades in 1 study. Likert quality grades for AICL were significantly higher than for *de novo* staff labels (mean: 6.2, SD: 0.9, $p = 0.001$) and edited staff labels (mean: 7.8, SD: 0.9, $p = 0.018$), and were higher but similar to *de novo* expert labels (mean: 8.1, SD: 0.8, $p = 0.30$). All *de novo* expert labels were also scored within the excellent range, whereas edited staff labels were scored in the moderate range for 2 studies, and *de novo* staff labels were scored in the moderate range in 13 studies. No scores were rated in the "poor" range. Of the six studies presented for discrimination between automated and human labeling, the blinded observer believed that three AI-only labels were annotated by the expert and believed that one attending segmentation was an AI-only inference label.

## Correlations: volumes, training cases, and time effort

Volumes ranged between 59.6 and 1,225.9 ml for hemothorax; between 38.6 and 1,062.6 ml for hemoperitoneum; and between 28.1 and 929.4 ml for pelvic hematoma. In the 57 patients, Pearson correlation (*r*) for i. *de novo* expert labeling time and ii. *de novo* staff labeling time vs. target volume were $r = 0.61$ (*de novo* expert), and 0.50 (*de novo* staff). Pearson's *r* for i. AICL and ii. staff label editing time vs. target volume were r = 0.36 (AICL), and 0.59 (staff edits). *de novo* and edited staff label times were moderately correlated with segmentation volume, but AICL times and volumes were only weakly correlated (40). Large AI-inferred volumes often required short editing times when much of the pathology was already correctly "filled-in", thus, AICL disassociates time effort from volume. This was not the case for

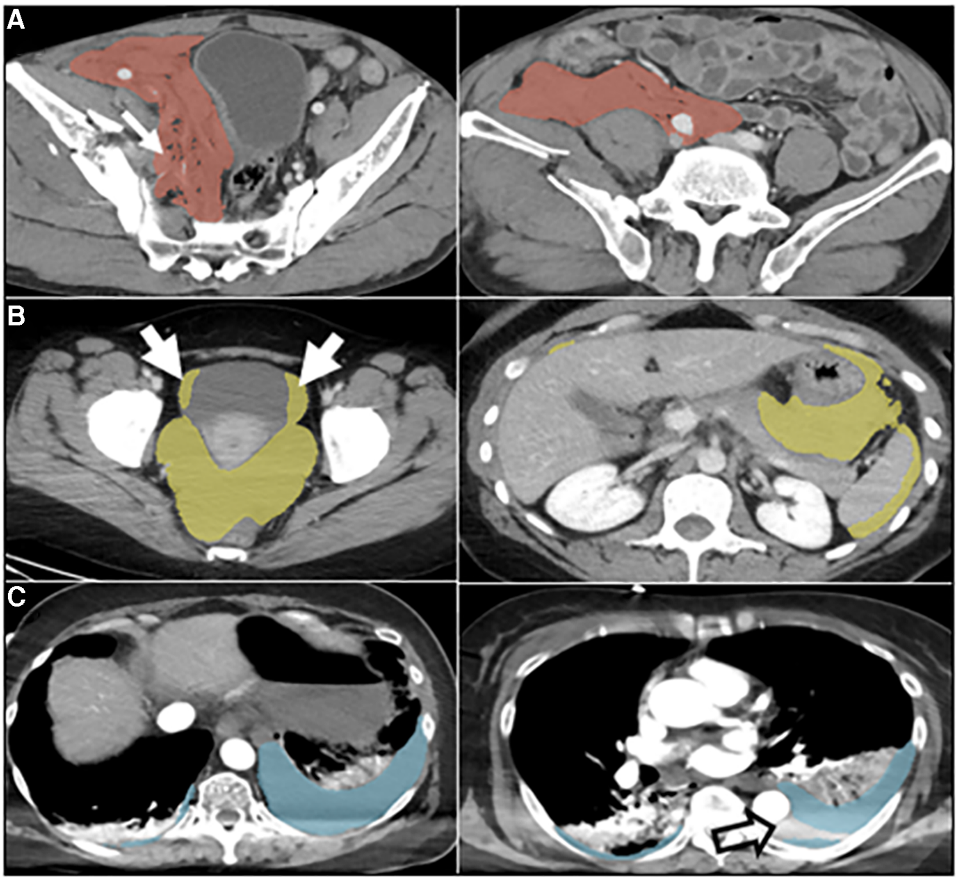

FIGURE 2
Automated labeling errors. Two slices are shown for each case example, one with an error and one without. (A) Pelvic hematoma (red). Arrow (left) shows extension of label into the right pelvic sidewall, requiring minimal edits. (B) Hemoperitoneum (yellow). Arrows show extension of label into the bladder margins, requiring minimal edits. (C) Hemothorax (blue). Open arrow, image right, shows unlabeled hemothorax from beam hardening artifact requiring more substantial editing. This artifact was not appreciated for hemoperitoneum or pelvic hematoma and may be related to the comparably small number of hemothorax training cases.

edited staff labels. Time effort to correct staff errors (which corresponded with significantly lower mean Likert scores) increased as targets increased in size.

Although correlation does not equal causation, we observed that the tasks with more training cases had higher AI label quality and decreasing AICL time. Mean AI-only Likert grades increased by task from 7.3 to 7.5 to 8.5, and median automated editing times decreased from 99 to 84 to 53 s for hemothorax (n = 77 training cases), hemoperitoneum ($n = 130$), and pelvic hematoma ($n = 253$) respectively. Hemothorax is considered the easiest task with well-defined crescentic shape and dependent distribution but had the fewest training examples, the longest editing times, and the lowest Likert scores.

## Discussion

A paucity of labeled data is generally acknowledged as the primary bottleneck for CAD research and development (6, 41). The cost, time, and human capital needed for voxelwise annotation of hemorrhage-related or similarly complex pathology on computed tomography is tremendous (42), and this is reflected in the comparatively small datasets reported in published studies to date (4). There is a recognized need for such tools, but few have come to market (4, 32).

A variety of emerging methods, including federated learning (43), synthetic data augmentation and semi-supervised approaches (30), active learning (13), and AI-assisted annotation or contour editing (7, 16, 17) are being explored for overcoming the scarcity of labeled data in cross-sectional imaging (6). In the meantime, investigators are beginning to use collaborative human-in-the-loop strategies with supervision and editing of automated labels to accelerate voxelwise annotation (19–21). In a recent COVID CT infiltrate segmentation challenge, automated labels edited by experts were used as the reference standard (19). In a pilot study with a maxillary sinus segmentation use case, nnU-net inference labels were edited and used for iterative retraining in several rounds, with final results compared to expert labels as the ground truth (21). Preliminary data from our early experience with deep learning-based pelvic hematoma segmentation showed that our model ignored outlier labeling

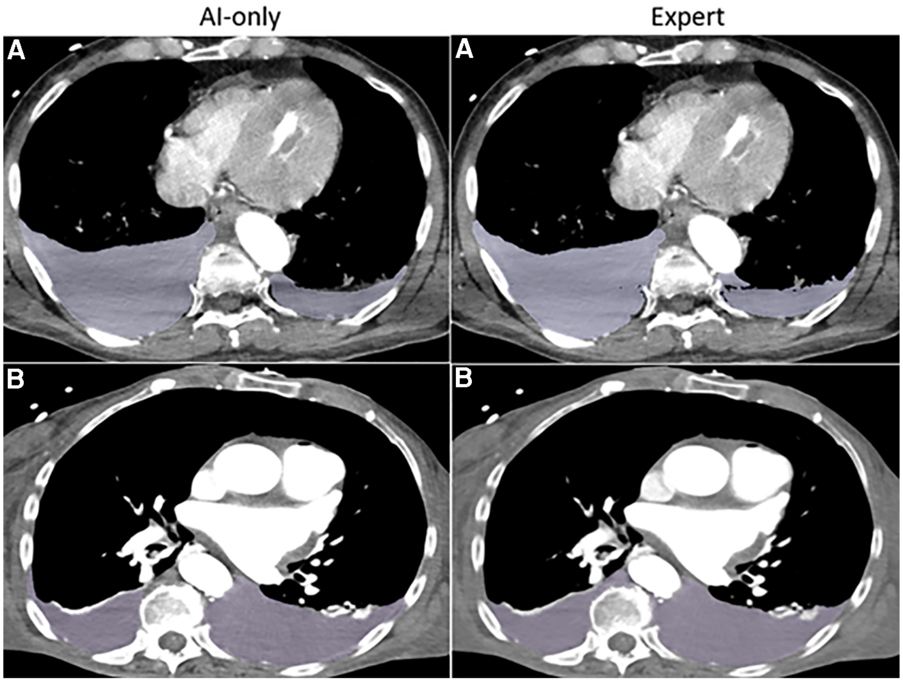

FIGURE 3A
Hemothorax. The independent observer correctly classified the hemothorax label source as *AI-only* or *expert* for case A, but not for case B. Both cases received scores in the 7–9 range and were considered adequate for clinical use.

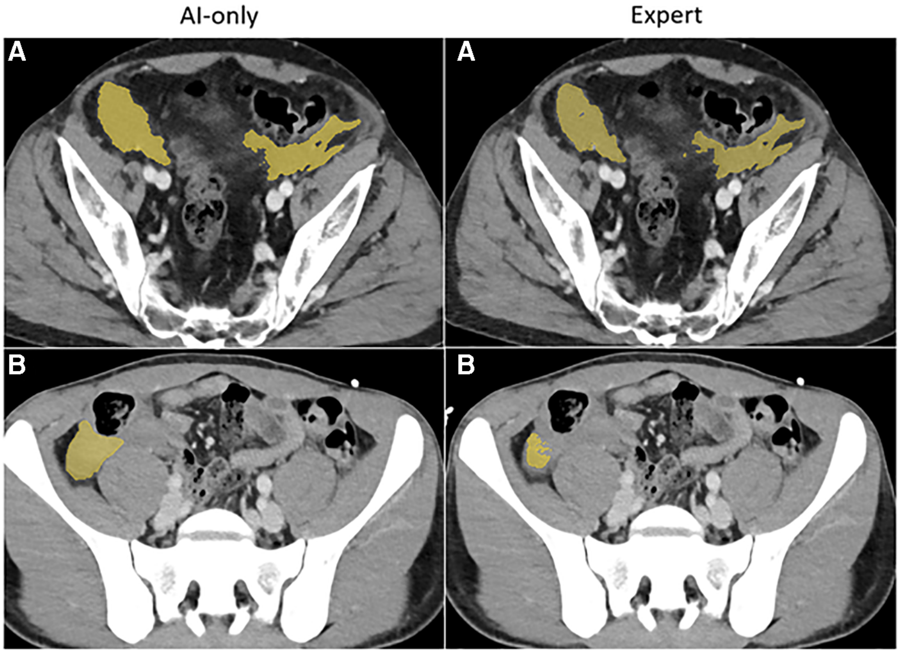

FIGURE 3B
Hemoperitoneum. The independent observer correctly classified the hemothorax label sources in A. Both AI-only and expert labels received scores in the excellent range. In case B, hemoperitoneum in the right paracolic gutter is incompletely labeled. The observer believed this to be due to algorithm error.

errors, suggesting that modest overlap metrics for this complex feature could be due in part to an imperfect reference standard. AICL was proposed as a means of overcoming both the time effort and fallibility associated with manual labeling for difficult tasks requiring high cognitive load. Such a strategy could potentially lead to higher agreement but lower quality

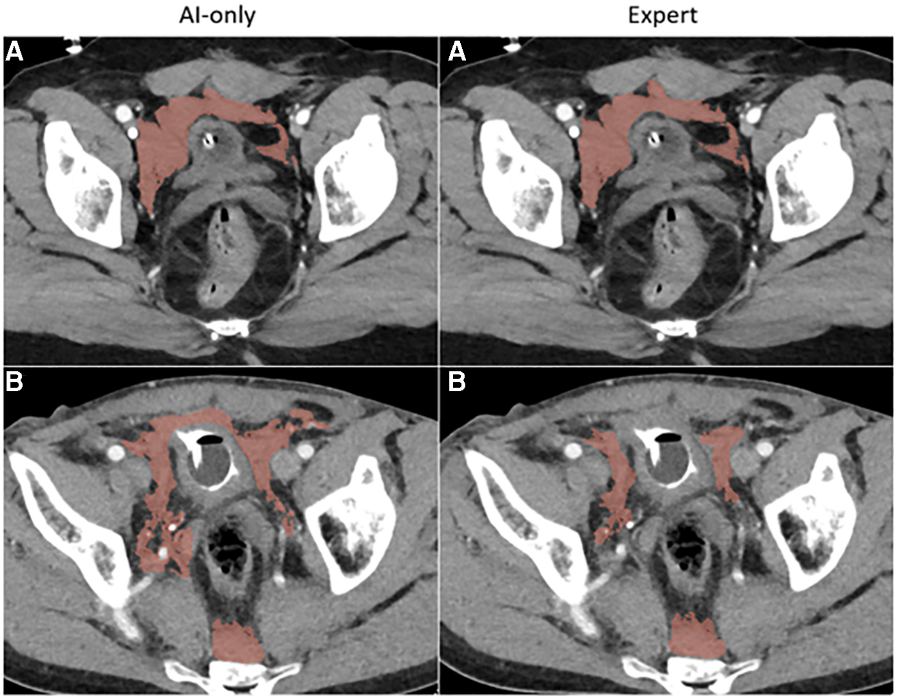

FIGURE 3C
Pelvic hematoma. The independent observer correctly classified case A, but not case B. The expert annotator under-segmented blood along the right pelvic sidewall. The independent observer interpreted this as machine error.

due to automation bias. To our knowledge, this had yet to be explored using semi-quantitative grading of label quality.

Labeling of Intracavitary pooled hemorrhage such as hemothorax, hemoperitoneum, and pelvic hematoma poses similar challenges to COVID infiltrates and surveillance of advanced malignancy; volumes can vary greatly, sometimes exceeding a liter of fluid for a given body cavity. Other challenges include the often multifocal, irregular, and ill-defined characteristics of pooled unopacified blood, and small differences in contrast with neighboring structures, including viscera and body wall (25, 29, 34).

CT is the routine imaging workhorse for patients with major trauma (44–46). Exsanguination remains a leading cause of preventable death, but clinical indices such as the shock index (SI) or Assessment of Blood Consumption (ABC) score are only modestly predictive and insensitive for life-threatening blood loss (47–49). The ability to predict outcomes or personalize treatment based on granular measurements of hemorrhage-related pathology on CT is a major precision medicine and personalized treatment-related value proposition of automated segmentation tools in the trauma domain (25, 30, 31, 34–36, 50–53), but large scale out-of-sample studies are required to confirm proof-of-principle.

In this work, we found that AICL for hemothorax, hemoperitoneum, and pelvic hematoma resulted in an overall 2.8-fold reduction in time effort over editing of staff labels, and an 8.7-fold reduction in time effort compared to *de novo* manual labeling by an expert. Further, AICL had significantly higher quality Likert scores (mean 8.4) than human *de novo* or edited labels (means from 6.3 to 7.8, $p = 0.02$ to $< 0.001$). For staff labels, increasing deviations in Likert grades from AICL grade as the reference standard correlated with lower DSC and longer editing times. The domain expert performing Likert grading was unable to correctly discriminate between manual expert labels and AI-only labels in the majority of patients, with an overall discordance rate of 66%.

Based on these results, the use of AICL appears justified as a reference standard for future experiments with our models and expert annotator. Such an approach can be extrapolated to very large datasets to substantially reduce costs and human capital as model-derived labels can be edited in place of those from trained staff. Our results do not necessarily generalize to other use cases or research teams and are not meant to justify using AICL as ground truth without appropriate quality control. However, our quality assessment framework could be used by other investigators and for other use cases to evaluate AICL annotation as an optimal high-throughput, high-quality reference standard. This would facilitate scaling voxelwise annotation to large cross-sectional imaging datasets for large-scale experiments. Once high-quality data annotation is complete and algorithms trained, they can be shadow-tested in the clinical environment with PACS-integrated software (54).

Before using AICL labels as ground truth, three conditions should ideally be met:

1. *AICL should take less time effort than human labeling or editing of human labels*. To achieve this, we recommend that sufficient manually-labeled seed data are used to generate at least moderate quality AI-only labels, using the proposed grading

scale. All grading should be performed by a blinded observer with relevant clinical domain expertise not involved in the labeling process. The number of manually labeled studies used as seed data will vary by task and will need to be determined on an individual basis.

2. *Semi-quantitative quality scores for AICL should be improved compared with human labels*. This ensures that automation or complacency bias does not adversely affect label quality.
3. *The external blinded observer performing Likert grading should have difficulty correctly discriminating AI and human labels.* This serves as an additional indicator that the AI-only labels are of sufficient quality to result in substantial time savings when using AICL. The proposed Likert grading system can also be employed as a preliminary assessment of the readiness of segmentation models for "in-the-wild" deployment.

Our study has several limitations. We were primarily constrained by the effort required from multiple participants in four different scenarios. Consequently, a small number of unseen studies were included. Models were trained using historical data from the same institution. Further, different numbers of patients were used to train each model. The pelvic hematoma model (trained with 253 CTs, compared to 130 for hemoperitoneum, and 77 for hemothorax), had the shortest editing times and highest Likert scores. Although we are unable to establish with certainty the degree to which this is related to more training cases, considering that pelvic hematoma is highly irregular and multi-compartmental, it is our experience that this represents a substantially more challenging task than labeling of hemothorax (compare for example **Figures 3A**,**C**). Based on comparative task simplicity alone, we would expect the hemothorax nnU-net model to yield the highest rather than the lowest quality labels. Additionally, our study was only powered to detect differences in time effort. However, differences in Likert scores were still found to be significant.

In the future, our approach could likely be combined with active learning and semi-automated AI-assisted annotation for further improvements in throughput.

## Conclusion

The precision-medicine-related value of volumetric imaging has been explored in a variety of domains such as surveillance for cancer progression, COVID infiltrate quantification, and hemorrhage burden in trauma. However, studies typically have insufficient sample sizes to ensure model generalizability and robustness in the clinical setting, which limits opportunities for productization and regulatory approval. The time effort and human capital required for manual voxelwise labeling prevents scaling to large datasets, particularly for pathology with irregular, multifocal, and potentially very large targets. Further, manual labels suffer from human error and are an imperfect ground truth.

We evaluated an AICL approach using models initialized on manually-labeled seed data. AICL annotations met three conditions for use as a reference standard: 1. They required significantly less time effort than human labeling, with time savings approaching an order of magnitude. 2. Semi-quantitative quality scores were improved compared with human-only labels, and 3. Automated and manual labels were not readily distinguishable. When these conditions are met, edited automated labels can be considered a high-quality gold standard for rapid curation of voxelwise data. Further improvements in throughput and performance may be possible with human-in-the-loop iterative retraining, interactive semi-automated labeling (AI-assisted annotation), and active learning strategies.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving human participants were reviewed and approved by University of Maryland IRB. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

DD, LZ, UB, MU, GL, NS, JH, HC, MK, UK, PC: Study concepts/study design or data acquisition or data analysis/interpretation; GL: Statistical analysis; DD: Original draft preparation: DD, LZ, UKB, MU, GL, NS, JH, HC, MK, UK, PC: Manuscript review and editing; DD, MU, LZ: Manuscript revision for important intellectual content; DD, LZ, UB, MU, GL, NS, JH, HC, MK, UK, PC: Manuscript final version approval; DD, MU: Literature research; DD: Funding acquisition. All authors contributed to the article and approved the submitted version.

## Funding

NIH NIBIB K08 EB027141-01A1

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors DD declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. *J Med Syst.* (2018) 42:1–13. doi: 10.1007/s10916-018-1088-1

2. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. *Radiology.* (2020) 295 (1):4–15. doi: 10.1148/radiol.2020192224

3. Heit J, Coelho H, Lima F, Granja M, Aghaebrahim A, Hanel R, et al. Automated cerebral hemorrhage detection using RAPID. *American Journal of Neuroradiology.* (2021) 42(2):273–8. doi: 10.3174/ajnr.A6926

4. Dreizin D, Staziaki PV, Khatri GD, Beckmann NM, Feng Z, Liang Y, et al. Artificial intelligence CAD tools in trauma imaging: a scoping review from the American society of emergency radiology (ASER) AI/ML expert panel. *Emerg Radiol.* (2023) 30(3):251–65. doi: 10.1007/s10140-023-02120-1

5. Ebrahimian S, Kalra MK, Agarwal S, Bizzo BC, Elkholy M, Wald C, et al. FDA-regulated AI algorithms: trends, strengths, and gaps of validation studies. *Acad Radiol.* (2022) 29(4):559–66. doi: 10.1016/j.acra.2021.09.002

6. Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, et al. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/the academy workshop. *Radiology.* (2019) 291 (3):781. doi: 10.1148/radiol.2019190613

7. Diaz-Pinto A, Mehta P, Alle S, Asad M, Brown R, Nath V, et al. *Deepedit: deep editable learning for interactive segmentation of 3D medical images. Data augmentation, labelling, and imperfections: second MICCAI workshop, DALI 2022, held in conjunction with MICCAI 2022*; September 22, 2022; Singapore: Proceedings: Springer (2022). p. 11–21

8. Diaz-Pinto A, Alle S, Ihsani A, Asad M, Nath V, Pérez-García F, et al. Monai label: A framework for ai-assisted interactive labeling of 3d medical images. arXiv preprint arXiv:220312362 (2022).

9. Hallitschke VJ, Schlumberger T, Kataliakos P, Marinov Z, Kim M, Heiliger L, et al. Multimodal Interactive Lung Lesion Segmentation: A Framework for Annotating PET/CT Images based on Physiological and Anatomical Cues. arXiv preprint arXiv:230109914 (2023).

10. McCrindle B, Zukotynski K, Doyle TE, Noseworthy MD. A radiology-focused review of predictive uncertainty for AI interpretability in computer-assisted segmentation. *Radiology: Artificial Intelligence.* (2021) 3(6):e210031. doi: 10.1148/ryai.2021210031

11. Iglesias JE, Konukoglu E, Montillo A, Tu Z, Criminisi A. *Combining generative and discriminative models for semantic segmentation of CT scans via active learning. Information processing in medical imaging: 22nd international conference, IPMI 2011*; July 3-8, 2011; Kloster Irsee, Germany: Proceedings 22: Springer (2011). p. 25–36

12. Pereira-Santos D, Prudêncio RBC, de Carvalho AC. Empirical investigation of active learning strategies. *Neurocomputing.* (2019) 326:15–27. doi: 10.1016/j.neucom.2017.05.105

13. Kuo W, Häne C, Yuh E, Mukherjee P, Malik J. *Cost-sensitive active learning for intracranial hemorrhage detection. Medical image computing and computer assisted intervention–MICCAI 2018. 21st International conference*; September 16-20, 2018; Granada, Spain: Proceedings, Part III 11: Springer (2018). p. 715–23

14. Sakinis T, Milletari F, Roth H, Korfiatis P, Kostandy P, Philbrick K, et al. Interactive segmentation of medical images through fully convolutional neural networks. arXiv preprint arXiv:190308205 (2019).

15. Wang G, Li W, Zuluaga MA, Pratt R, Patel PA, Aertsen M, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on Medical Imaging.* (2018) 37(7):1562–73. doi: 10.1109/TMI.2018.2791721

16. Berg S, Kutra D, Kroeger T, Straehle CN, Kausler BX, Haubold C, et al. Ilastik: interactive machine learning for (bio) image analysis. *Nat Methods.* (2019) 16 (12):1226–32. doi: 10.1038/s41592-019-0582-9

17. Bai T, Balagopal A, Dohopolski M, Morgan HE, McBeth R, Tan J, et al. A proof-of-concept study of artificial intelligence–assisted contour editing. *Radiology: Artificial Intelligence.* (2022) 4(5):e210214. doi: 10.1148/ryai.210214

18. Amrehn M, Gaube S, Unberath M, Schebesch F, Horz T, Strumia M, et al. UI-Net: Interactive artificial neural networks for iterative image segmentation based on a user model. arXiv preprint arXiv:170903450 (2017).

19. Roth HR, Xu Z, Tor-Díez C, Jacob RS, Zember J, Molto J, et al. Rapid artificial intelligence solutions in a pandemic—the COVID-19-20 lung CT lesion segmentation challenge. *Med Image Anal.* (2022) 82:102605. doi: 10.1016/j.media.2022.102605

20. Kim T, Lee KH, Ham S, Park B, Lee S, Hong D, et al. Active learning for accuracy enhancement of semantic segmentation with CNN-corrected label curations: evaluation on kidney segmentation in abdominal CT. *Sci Rep.* (2020) 10 (1):366. doi: 10.1038/s41598-019-57242-9

21. Jung S-K, Lim H-K, Lee S, Cho Y, Song I-S. Deep active learning for automatic segmentation of maxillary sinus lesions using a convolutional neural network. *Diagnostics.* (2021) 11(4):688. doi: 10.3390/diagnostics11040688

22. Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation:"data awareness" that matters. *Ann Transl Med.* (2020) 8(7):501. doi: 10.21037/atm.2020.03.63

23. Rädsch T, Reinke A, Weru V, Tizabi MD, Schreck N, Kavur AE, et al. Labelling instructions matter in biomedical image analysis. *Nat Mach Intell.* (2023) 5:273–83. doi: 10.1038/s42256-023-00625-5

24. Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The medical segmentation decathlon. *Nat Commun.* (2022) 13(1):4128. doi: 10.1038/s41467-022-30695-9

25. Dreizin D, Zhou Y, Fu S, Wang Y, Li G, Champ K, et al. A multiscale deep learning method for quantitative visualization of traumatic hemoperitoneum at CT: assessment of feasibility and comparison with subjective categorical estimation. *Radiology.* (2020) 2(6):e190220. doi: 10.1148/ryai.2020190220

26. Harris RJ, Kim S, Lohr J, Towey S, Velichkovich Z, Kabachenko T, et al. Classification of aortic dissection and rupture on post-contrast CT images using a convolutional neural network. *J Digit Imaging.* (2019) 32(6):939–46. doi: 10.1007/s10278-019-00281-5

27. Choi J, Mavrommati K, Li NY, Patil A, Chen K, Hindin DI, et al. Scalable deep learning algorithm to compute percent pulmonary contusion among patients with rib fractures. *J Trauma Acute Care Surg.* (2022) 93(4):461–6. doi: 10.1097/TA.0000000000003619

28. Röhrich S, Hofmanninger J, Negrin L, Langs G, Prosch H. Radiomics score predicts acute respiratory distress syndrome based on the initial CT scan after trauma. *Eur Radiol.* (2021) 31:5443–53. doi: 10.1007/s00330-020-07635-6

29. Dreizin D, Zhou Y, Zhang Y, Tirada N, Yuille AL. Performance of a deep learning algorithm for automated segmentation and quantification of traumatic pelvic hematomas on CT. *J Digit Imaging.* (2020) 33(1):243–51. doi: 10.1007/s10278-019-00207-1

30. Zhou Y, Dreizin D, Wang Y, Liu F, Shen W, Yuille AL. External attention assisted multi-phase splenic vascular injury segmentation with limited data. *IEEE Trans Med Imaging.* (2021) 41(6):1346–57. doi: 10.1109/TMI.2021.3139637

31. Zhou Y, Dreizin D, Li Y, Zhang Z, Wang Y, Yuille A. *Multi-scale attentional network for multi-focal segmentation of active bleed after pelvic fractures. Machine learning in medical imaging. 10th International workshop, MLMI 2019, held in conjunction with MICCAI 2019*; October 13, 2019; Shenzhen, China: Proceedings 10: Springer (2019). p. 461–9

32. Agrawal A, Khatri GD, Khurana B, Sodickson AD, Liang Y, Dreizin D. A survey of ASER members on artificial intelligence in emergency radiology: trends, perceptions, and expectations. *Emerg Radiol.* (2023) 30(3):267–77. doi: 10.1007/s10140-023-02121-0

33. Lee S, Summers RM. Clinical artificial intelligence applications in radiology: chest and abdomen. *Radiol Clin North Am.* (2021) 59(6):987–1002. doi: 10.1016/j.rcl.2021.07.001

34. Dreizin D, Nixon B, Hu J, Albert B, Yan C, Yang G, et al. A pilot study of deep learning-based CT volumetry for traumatic hemothorax. *Emerg Radiol.* (2022) 29 (6):995–1002. doi: 10.1007/s10140-022-02087-5

35. Dreizin D, Bodanapally U, Boscak A, Tirada N, Issa G, Nascone JW, et al. CT Prediction model for major arterial injury after blunt pelvic ring disruption. *Radiology.* (2018) 287(3):1061–9. doi: 10.1148/radiol.2018170997

36. Dreizin D, Champ K, Dattwyler M, Bodanapally U, Smith EB, Li G, et al. Blunt splenic injury in adults: association between volumetric quantitative CT parameters and intervention. *J Trauma Acute Care Surg.* (2023) 94(1):125–32. doi: 10.1097/TA.0000000000003684

37. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* (2021) 18(2):203–11. doi: 10.1038/s41592-020-01008-z

38. Nixon B, Albert A, Yan C, Yang G, Unberath M, Jeudy J, et al. *Computational modeling of hemorrhage-related outcomes after chest trauma*

*using deep learning-based CT hemothorax volumetry and point-of-care clinical parameters radiologic society of North America annual meeting and scientific assembly*; Chicago, IL (2021).

39. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. *J Exp Psychol*. (2012) 141(1):2. doi: 10.1037/a0024338

40. Dancey CP, Reidy J. *Statistics without maths for psychology*. London: Pearson (2017). p. 632.

41. Allen B Jr, Seltzer SE, Langlotz CP, Dreyer KP, Summers RM, Petrick N, et al. A road map for translational research on artificial intelligence in medical imaging: from the 2018 national institutes of health/RSNA/ACR/the academy workshop. *J Am Coll Radiol*. (2019) 16(9):1179–89. doi: 10.1016/j.jacr.2019.04.014

42. Dreizin D, Bodanapally UK, Neerchal N, Tirada N, Patlas M, Herskovits E. Volumetric analysis of pelvic hematomas after blunt trauma using semi-automated seeded region growing segmentation: a method validation study. *Abdom Radiol*. (2016) 41(11):2203–8. doi: 10.1007/s00261-016-0822-8

43. Yan R, Qu L, Wei Q, Huang S-C, Shen L, Rubin D, et al. Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging. *IEEE Trans Med Imaging*. (2023) 42(7):1932–43. doi: 10.1109/TMI.2022.3233574

44. Dreizin D, Munera F. Blunt polytrauma: evaluation with 64-section whole-body CT angiography. *RadioGraphics*. (2012) 32(3):609–31. doi: 10.1148/rg.323115099

45. Dreizin D, Munera F. Multidetector CT for penetrating torso trauma: state of the art. *Radiology*. (2015) 277(2):338–55. doi: 10.1148/radiol.2015142282

46. Costantini TW, Coimbra R, Holcomb JB, Podbielski JM, Catalano RD, Blackburn A, et al. Pelvic fracture pattern predicts the need for hemorrhage control intervention—results of an AAST multi-institutional study. *J Trauma Acute Care Surg*. (2017) 82(6):1030–8. doi: 10.1097/TA.0000000000001465

47. Schroll R, Swift D, Tatum D, Couch S, Heaney JB, Llado-Farrulla M, et al. Accuracy of shock index versus ABC score to predict need for massive transfusion in trauma patients. *Injury*. (2018) 49(1):15–9. doi: 10.1016/j.injury.2017.09.015

48. Rau C-S, Wu S-C, Kuo SC, Pao-Jen K, Shiun-Yuan H, Chen Y-C, et al. Prediction of massive transfusion in trauma patients with shock index, modified shock index, and age shock index. *Int J Environ Res Public Health*. (2016) 13(7):683. doi: 10.3390/ijerph13070683

49. Olaussen A, Blackburn T, Mitra B. Fitzgerald M. Shock index for prediction of critical bleeding post-trauma: a systematic review. *Emerg Med Australas*. (2014) 26 (3):223–8. doi: 10.1111/1742-6723.12232

50. Dreizin D, Zhou Y, Chen T, Li G, Yuille AL, McLenithan A, et al. Deep learning-based quantitative visualization and measurement of extraperitoneal hematoma volumes in patients with pelvic fractures: potential role in personalized forecasting and decision support. *J Trauma Acute Care Surg*. (2020) 88(3):425–33. doi: 10.1097/TA.0000000000002566

51. Dreizin D, Yu T, Motley K, Li G, Morrison JJ, Liang Y. Blunt splenic injury: assessment of follow-up CT utility using quantitative volumetry. *Front Radiol*. (2022) doi: 10.3389/fradi.2022.941863. Epub 2022 Jul 22

52. Dreizin D, Chen T, Liang Y, Zhou Y, Paes F, Wang Y, et al. Added value of deep learning-based liver parenchymal CT volumetry for predicting major arterial injury after blunt hepatic trauma: a decision tree analysis. *Abdom Radiol*. (2021) 46 (6):2556–66. doi: 10.1007/s00261-020-02892-x

53. Chen H, Unberath M, Dreizin D. Toward automated interpretable AAST grading for blunt splenic injury. *Emerg Radiol*. (2023) 30(1):41–50. doi: 10.1007/s10140-022-02099-1

54. Zhang L, LaBelle W, Unberath M, Chen H, Hu J, Li G, et al. A vendor-agnostic, PACS integrated, and DICOM-compatible software-server pipeline for testing segmentation algorithms within the clinical radiology workflow. (2023).