nature communications

Article

https://doi.org/10.1038/s41467-025-67581-z

# Multicenter study on the versatility and adoption of AI-driven automated radiotherapy planning across cancer types

Received: 16 June 2025

Accepted: 2 December 2025

Published online: 15 December 2025

Check for updates

Lei Yu[1,2,3,4,11], Qianxi Ni[5,6,11], Binbing Wang[7,8,11], Kang Zhang[9], Feng Shi[10], Shixiong Huang[5,6], Guoping Shan[7,8], Yang Zhong[1,2,3,4], Ying Guo[1,2,3,4], Zhen Zhang[1,2,3,4] ✉, Jiazhou Wang[1,2,3,4] ✉ & Weigang Hu[1,2,3,4] ✉

Deep learning (DL) -based automated treatment planning (ATP) shows significant promise in streamlining radiotherapy workflow and reducing variability in plan quality. However, it often lacks the flexibility needed for achieving individualized trade-offs in real-world practice. Herein, we propose a hybrid strategy by integrating DL-based dose prediction with clinical-goal-guided inverse optimization to generate directly deliverable plans within five minutes. DL models for five disease sites were trained separately using datasets from a single institution and were tested retrospectively for clinical application among three institutions, with tailored prioritized clinical goals. We find that over 80% of the 250 auto-plans met clinical criteria, and 60% were preferred over manual plans in blinded reviews. Dosimetric analyses show that the auto-plans quantitatively matched or exceeded the quality of human-driven plans. This study highlights ATP's potential to transform radiotherapy practice, with ongoing efforts aimed at refining its versatility and adoption across diverse clinical settings.

Radiotherapy planning involves designing an optimal dose distribution that ensures adequate coverage of planning target volume (PTV) with minimizing exposure to organs at risk (OARs). This process, vital to treatment efficacy, has become more sophisticated since the introduction of intensity-modulated radiotherapy (IMRT) in the 1980s. Efforts to improve plan quality have increased the time, cost, and complexity of manual planning[1]. Human-driven plans, shaped by trial and error, depend heavily on planner's expertise. Variability in hyperparameter tuning, such as constraint weights and dose thresholds, often leads to inconsistencies in plan quality, which may ultimately undermine patient outcomes. In fact, if better target conformality and/or homogeneity, as well as improved normal tissue sparing should have been achieved at no cost, patients will be exposed to substantial excess risk of local failure or complications due to suboptimal planning[2].

To improve efficiency and reduce variability in plan quality, automated radiotherapy treatment planning (ATP) has been extensively studied[3]. Several commercial treatment planning systems (TPSs) now offer mature ATP solutions, including protocol-based optimization (e.g. AutoPlanning™, Philips Radiation Oncology Systems, Fitchburg, WI), which mimics human-driven iterative adjustments, knowledge-based planning (e.g. RapidPlan™, Varian Medical

[1]Department of Radiation Oncology, Fudan University Shanghai Cancer Center, Shanghai, China. [2]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China. [3]Shanghai Clinical Research Center for Radiation Oncology, Shanghai, China. [4]Shanghai Key Laboratory of Radiation Oncology, Shanghai, China. [5]The Affiliated Cancer Hospital of Xiangya School of Medicine, Central South University, Changsha, Hunan, China. [6]Hunan Cancer Hospital, Changsha, Hunan, China. [7]Zhejiang Cancer Hospital, Hangzhou, Zhejiang, China. [8]Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, Zhejiang, China. [9]Radiotherapy Business Unit, Shanghai United Imaging Healthcare Co. Ltd., Shanghai, China. [10]Department of Research and Development, Shanghai United Imaging Intelligence Co. Ltd., Shanghai, China. [11]These authors contributed equally: Lei Yu, Qianxi Ni, Binbing Wang. ✉e-mail: zhen_zhang@fudan.edu.cn; wjiazhou@gmail.com; jackhuwg@gmail.com

Systems, Palo Alto, CA), which leverages prior clinical plan database, and multi-criteria optimization (e.g. RayStation multi-criteria optimization (MCO), RaySearch Laboratories, Sweden), which utilizes Pareto optimality. These approaches are now widely adopted in clinical practice, with numerous studies[4–8] confirming ATP's efficiency and its noninferiority to manual planning. Furthermore, recent advancements of deep learning (DL) in ATP have led to significant improvements in both the accuracy and efficiency of dose prediction and final plan generation[9,10], marking a new era in ATP development.

For AI-based applications like ATP, a frequently asked question is: how can it best serve clinical practice? That is to say, should the goal be to surpass human planners entirely, as AlphaGo did[11], or to efficiently generate clinically usable plans with minimal human involvement. The answer lies in three critical factors. First, ATP plans must meet essential clinical constraints, ensuring safety for patient treatment. These constraints, defined by the Radiation Therapy Oncology Group (RTOG) protocols, represent "clinical acceptability" and can be explicitly incorporated into algorithms. Second, plans should strive for Pareto optimality—maximizing OARs sparing without sacrificing tumor control—known as "clinical optimality". Lastly, efficiency is a key concern, particularly for on-couch treatments. This includes rapid planning, minimal human involvement, ease of adjustment, and seamless workflow integration, collectively termed "clinical applicability". In current ATP practice, achieving clinical acceptability is non-negotiable, while optimality and applicability are desirable bonuses over manual planning. However, balancing these factors still remains a significant challenge[3,12–14].

This study seeks to address these challenges through a retrospective analysis of ATP implementation across multiple institutions and disease sites, using a single-institutional DL framework. Instead of merely comparing auto-plans to manual plans, we focus on evaluating their acceptability and generalizability, as well as improving their applicability in real-world settings. This includes accounting for variations in target definitions, beam arrangements, and evaluation protocols across institutions. Figure 1 outlines the study framework.

Inspired by knowledge-based planning (KBP) dose prediction and MCO wish lists, this study introduces a hybrid ATP method that integrates DL-based dose prediction with clinical-goal-guided inverse optimization, enabling the generation of directly deliverable plans in a treatment planning system (uTPS, United Imaging Healthcare, Shanghai, China). In this system, we employed a channel attention densely connected U-Net (CAD-UNet)[15] to predict voxel-based dose distributions using patient-specific CT scans and contours as input (Supplementary Fig.1 and 2). Models were trained and validated separately for five disease sites—nasopharyngeal carcinoma (NPC), lung, breast, cervix, and rectum—using high-quality datasets from a single institution (Institution A) (Supplementary Table 1). For each site, a clinical goal list was prioritized to balance target coverage and OARs sparing, providing a reference for DVH prediction (Supplementary Tables 2 and 3). A tolerance parameter was used to correct bias in each predicted goal (Supplementary Tables 3 and 4). Executable plans were generated in uTPS through inverse optimization guided by these clinical goals, typically within 5 min. The prediction models developed by Institution A have been implemented in the ATP module of the uTPS and are available for clinical use with permission.

The acceptability evaluation phase tested the ATP solution retrospectively across three institutions: internal institution A and two external institutions B and C. Single-institutional DL models and fixed clinical goal lists were used, with tolerance adjustments made for Institution C's local protocols (Fig. 1 and Supplementary Tables 3 and 4). Each site enrolled 30, 10, and 10 patients, randomly selected in A, B, and C, respectively. The enrolled cases ranged in target definitions, delivery techniques, and prescribed dosages, as shown in Table 1.

ATP plans (250 cases in total) were independently generated, maintaining original beam configurations and without patient-specific goal adjustments. These plans were compared to clinically approved manual plans (MNL plans) based on subjective and objective criteria. Blinded assessments were conducted by three local physicians per site, who evaluated the plans for clinical acceptability, dosimetric preference, and generation method (ATP or MNL). Feedback from these assessments was collected for further analysis. Dosimetric performance was evaluated using RTOG-defined DVH endpoints[16–20], with statistical significance determined through paired two-tailed $t$-tests ($p < 0.05$). Dosimetric parameters between ATP-preferred and MNL-preferred cohorts were compared to quantify the strengths and weaknesses of the ATP solution. Differences across institutions in planning protocols and assessment results were examined to evaluate the solution's generalizability and robustness in varying clinical contexts.

In the applicability improvement phase, particular focus was placed on the auto-plans deemed unacceptable or unfavorable during the evaluation phase. Tailored strategies were then implemented

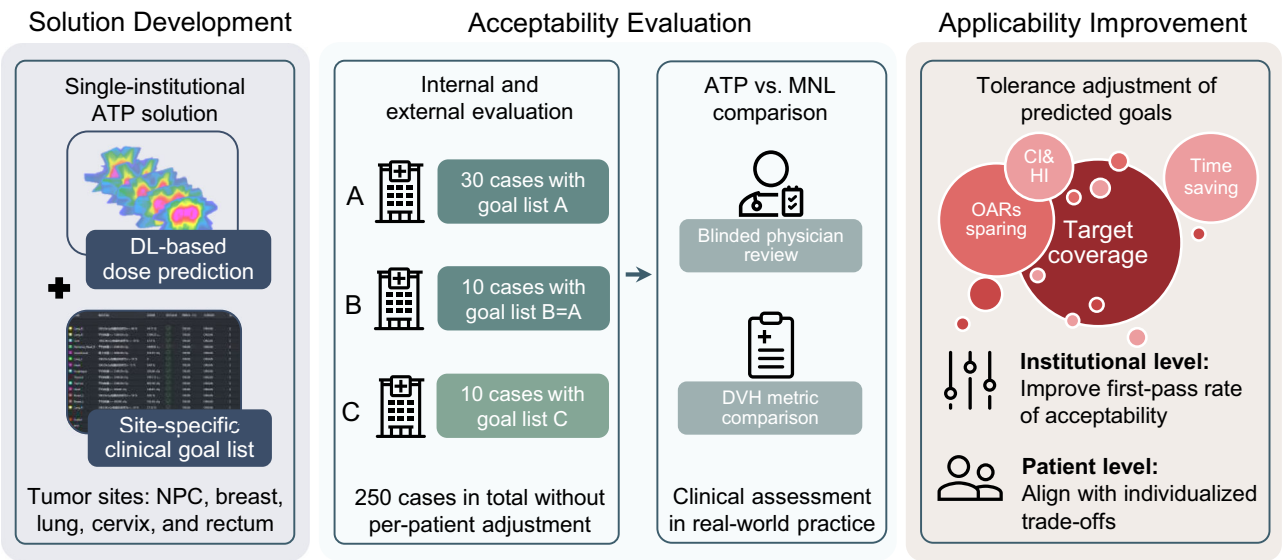


**Fig. 1 | Framework of the current study on the proposed ATP solution.** The third-party elements are from Health icons (https://healthicons.org/).

**Table 1 | Characteristics of the enrolled patient cohorts across different institutions**

| Site | Characteristics | Institution A (30) | Institution B (10) | Institution C (10) |
|---|---|---|---|---|
| NPC | PTV volume (ml) | 779.69 ± 132.27, (554.50 ~ 996.62) | 764.09 ± 223.12, (497.27 ~ 1074.04) | 352.87 ± 96.81, (199.91 ~ 565.87) |
| | Treatment technique | 9/11-field dIMRT (30) | 2/3-arc VMAT (10) | 2/4-arc VMAT (10) |
| | Prescribed dose | 66-60-54 Gy/30 F (8)<br>70.4-66-60-54 Gy/32 F (15)<br>70-66-63-56 Gy/35 F (7) | 70-66-60-56 Gy/35 F (10) | 69.96-66-61.05-54.45 Gy/33 F (7)<br>69.96-62.7-60.06-54.45 Gy/33 F (3) |
| Lung | PTV volume (ml) | 458.33 ± 309.09, (49.38 ~ 1353.09) | 226.44 ± 110.19, (82.31 ~ 413.56) | 251.86 ± 76.45, (142.09 ~ 414.88) |
| | Treatment technique | 7 ~ 10-field sIMRT (3)<br>7 ~ 10-field dIMRT (27) | 2-arc VMAT (10) | 4/6-arc VMAT (10) |
| | Prescribed dose | 50 Gy/25 F (20)<br>60 Gy/30 F (10) | 60 Gy/30 F (10) | 60 Gy/30 F (10) |
| Breast | PTV volume (ml) | 697.18 ± 308.69, (81.59 ~ 1310.56) | 774.43 ± 292.49, (385.22 ~ 1220.02) | 749.93 ± 156.63, (488.47 ~ 1004.97) |
| | Treatment technique | 8/9-field sIMRT (4)<br>8/9-field dIMRT (13)<br>4-arc VMAT (13) | 2/4-arc VMAT (10) | 4/6-arc VMAT (10) |
| | Prescribed dose | 50 Gy/25 F (13)<br>42.56 Gy/16 F (6)<br>40.05 Gy/15 F (11) | 50 Gy/25 F (10) | 43.5 Gy/15 F (1)<br>50 Gy/25 F (9) |
| Rectum | PTV volume (ml) | 1120.63 ± 205.57, (782.56 ~ 1602.94) | 1062.01 ± 224.82, (729.56 ~ 1468.91) | 788.45 ± 145.28, (572.03 ~ 1085.41) |
| | Treatment technique | 9-field sIMRT (10)<br>9-field dIMRT (3)<br>2-arc VMAT (17) | 2-arc VMAT (10) | 2-arc VMAT (10) |
| | Prescribed dose | 50 Gy/25 F (22)<br>25 Gy/5 F (8) | 50-45 Gy/25 F (10) | 25 Gy/5 F (10) |
| Cervix | PTV volume (ml) | 1470.85 ± 173.03, (1205.56 ~ 1892.88) | 1460.91 ± 237.75, (1027.88 ~ 1767.72) | 1120.45 ± 364.49, (352.66 ~ 1709.44) |
| | Treatment technique | 9-field dIMRT (30) | 2-arc VMAT (10) | 2-arc VMAT (10) |
| | Prescribed dose | 50.4 Gy/28 F (13)<br>58.8-50.4 Gy/28 F (17) | 45 Gy/25 F (10) | 45 Gy/25 F (10) |

The numbers in bracket indicate either the range of PTV volume or the respective number of patients corresponding to each treatment technique or prescription regimen.

within the ATP framework to address clinical concerns and improve plan quality. In addition, total time savings of the end-to-end ATP process were also analyzed in real-world settings.

In this work, the proposed ATP solution demonstrates broad acceptability and strong physician preference across multiple institutions and cancer types. Compared with manual planning, ATP reduces the total time required to finalize a deliverable plan in real-world scenarios by approximately 40%. This ATP approach is anticipated to enable immediate availability and enhanced applicability in diverse clinical contexts, while also providing a practical methodology to transition ATP from single-institutional validation to multi-institutional deployment.

## Results

### Subjective assessment

Visual comparisons of dose distributions between ATP and MNL plans across five sites are presented in Fig. 2, and the subjective assessment results are summarized in Fig. 3. Across all institutions, 82% of ATP plans (205/250) were unanimously deemed clinically acceptable by three reviewing physicians, and 60% (149/250) were preferred over MNL plans. The highest clinical acceptability rate was observed for NPC cases, achieving 100% consensus across all institutions. Moderate acceptability rates were noted for breast (84%), cervix (80%), and rectum (80%) cases, while lung cases had a lower rate of 66%. Preference for ATP plans over MNL plans also varied by site. Cervix cases had the highest preference rate (76%, 38/50), while breast cases had the lowest (46%, 23/50). NPC, lung, and rectum cases showed preference rates of 54%, 60%, and 62%, respectively. On average, 61% of ATP plans were recognized as AI-generated, with recognition rates ranging from 40% to 83%, showing no clear dependence on site or institution. Physicians noted that plan sources were nearly indistinguishable, often leading to arbitrary choices in some cases.

Inter-institutional comparisons reveal notable differences in ATP performance. Institutions A and C exhibited similar ATP performance, with Institution C achieving a 100% acceptability rate, reflecting the robust generalization of the method with tailored goal lists. In contrast, Institution B showed significant variability, with acceptability and preference rates dropping as low as 30% for breast and lung cases. Interestingly, NPC and cervix cases performed better in Institution B and C compared to Institution A, both in terms of acceptability and preference. These findings highlight the variability in ATP performance across institutions and the importance of site-specific and institution-specific factors in clinical implementation.

This analysis also emphasizes inter-observer variability in assessing ATP's effectiveness and suitability. As shown in Fig. 3, physicians at Institution C exhibited greater agreement compared to those at Institutions A and B. When majority consensus among reviewers was considered, the overall acceptance rate for ATP plans rose to 92%, with 72% of plans being preferred. More than 60% of ATP plans were favored over MNL plans across all sites. Reviewers' comments underscored the importance of target coverage, conformality, homogeneity, and OARs sparing (Supplementary Fig. 3). However, variability arose from differing perspectives on how to balance these factors, even when the plans met clinical acceptance criteria, contributing to discrepancies in final decisions.

### Objective comparison

To quantify the differences between ATP and manual planning, plan quality metrics were compared across institutions and preferable cohorts. Dot plots in Fig. 4 depict the average differences in dosimetric parameters of ATP plans relative to MNL plans across five sites. The results show that ATP plans generally achieved comparable or superior performance in conformality indices (CI), homogeneity indices (HI) of target and sparing of most OARs, while maintaining target coverage and similar beam modulation (total MUs) to MNL plans. Compromises

https://doi.org/10.1038/s41467-025-67581-z

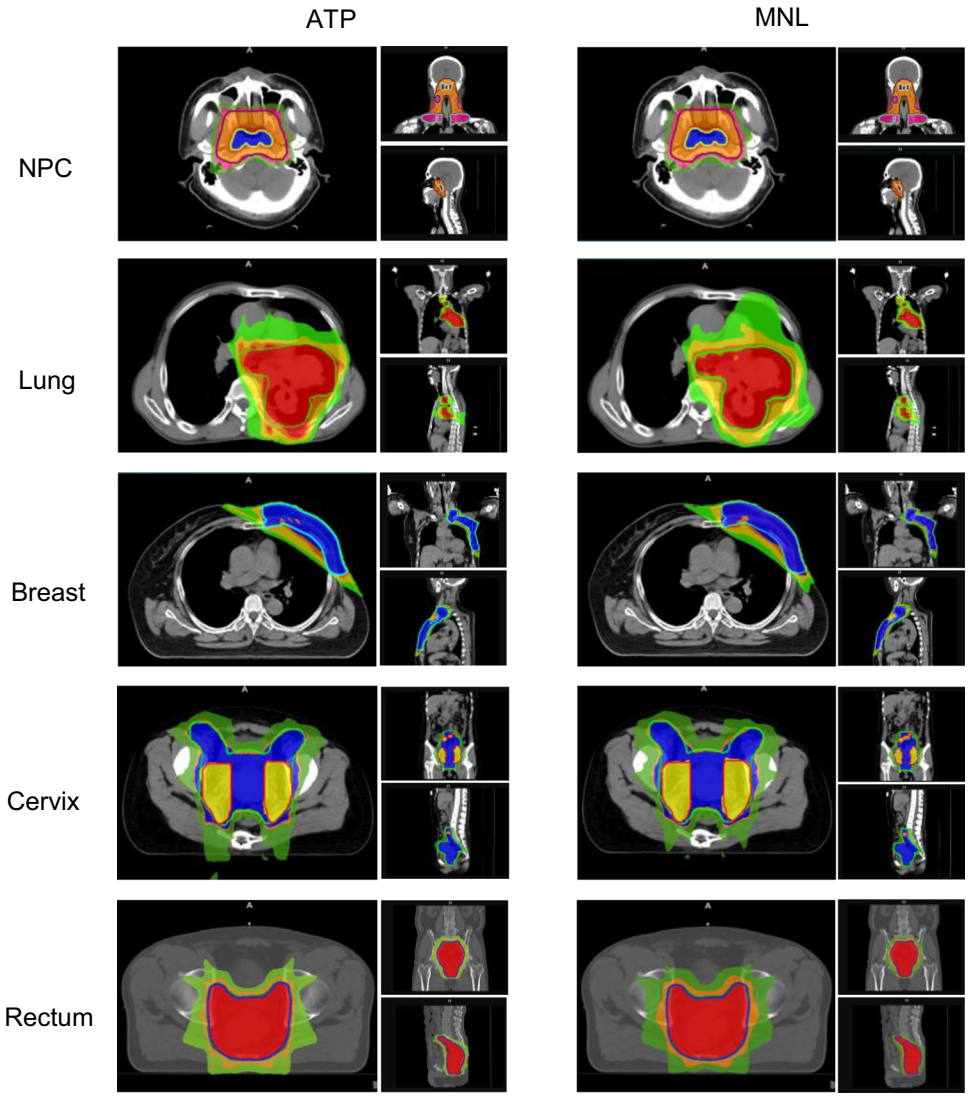


**Fig. 2 | Visual comparisons of representative dose distributions between automated (ATP) plans and manual (MNL) plans across five disease sites.** The sectional visualizations of each disease site are derived from an example case in the retrospective evaluation, illustrating the dose distribution patterns of both ATP and MNL plans side by side. The color-coded filled regions represent the dose levels of 100%, 95%, 80%, and 60% of the prescription dose.

in sparing of individual OARs were observed in certain locations, yet these had negligible statistical significance. The site-specific analyses are described as follows.

- For NPC cases, the ATP plans showed better sparing of brainstem, TMJs, oral cavity at Institution A, and improved overall target conformality at Institution B, whereas exhibited slightly deviated but comparable dosimetric parameters to the MNL plans at Institution C.
- For breast cases, discrepancies in comparing results were observed among the institutions. While lower V5Gy in ipsilateral and contralateral lungs as well as reduced mean dose of ipsilateral humeral head were achieved at Institution A, the ATP plans resulted in moderate sparing of ipsilateral lung at Institution B and C, but significant over-sparing of ipsilateral humeral head in B and contralateral breast in C. Relatively speaking, the ATP plans in Institution C demonstrated more balanced improvements in contrast to MNL plans (except for PTV HI and V20Gy of lung), owing to tailored goals to local guidelines. This fact was also reflected by higher reviewer preference in C than that in B.
- In the lung cohort, the results of plan quality comparison were less differentiating, where slightly better lung sparing in A and improved target conformality in B were noticed. There seemed no sufficient dosimetric merit/demerit to explain the wide difference in subjective selection. In Institution C, no significant dosimetric advantage of ATP was observed, despite a higher selection rate (70%) compared to Institution B (30%). At Institution A, despite the improvement in lung sparing, ATP plans were preferred in only 67% of cases. Even when MNL plans were favored, the counterpart auto-plans were quantitatively noninferior across all metrics, suggesting that subjective selection may have been influenced by minor perceived flaws.
- In spite of marginal weaknesses of PTV V95% in A and bladder V45Gy in B for cervical cases, ATP plans in the pelvic region generally showed significant improvements in target conformality, homogeneity, and substantial OARs sparing across institutions, leading to notable reviewer recognition.

The superior plan quality of the ATP preferable cohort, consistent with subjective selections, underscores its strengths in various contexts. Importantly, no significant deficiencies were observed in unselected ATP plans across the five sites, with some exhibiting better

https://doi.org/10.1038/s41467-025-67581-z

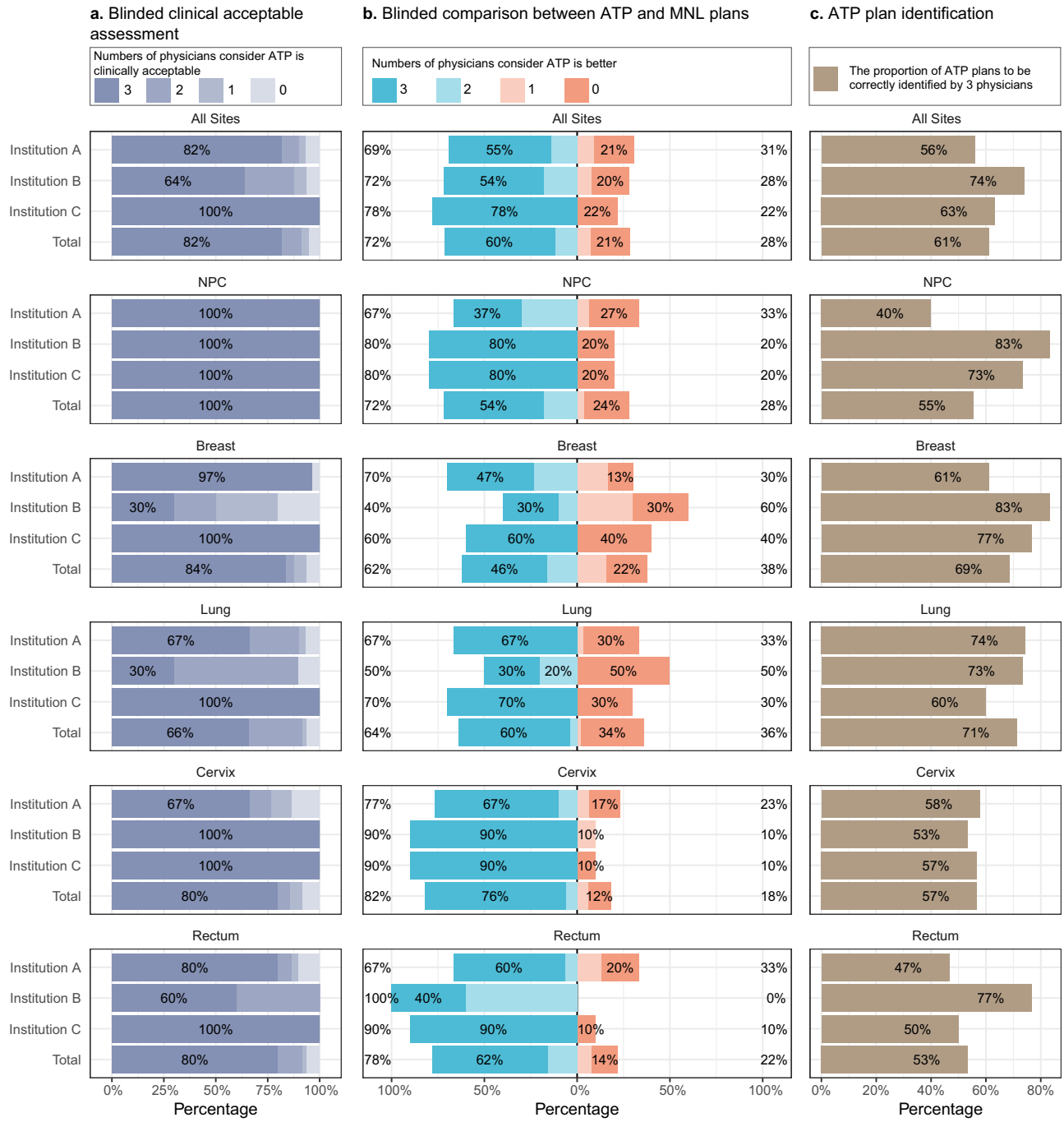


**Fig. 3 | Results of blinded physician assessments across all sites and institutions.** For each site, $n = 30, 10, 10$ cases at Institution A, B, C, respectively. **a** and **b** show the percentages of the ATP plans considered to be clinically acceptable and preferable to MNL plans by different numbers of physicians, respectively. **c** shows the average proportion of the ATP plans to be correctly identified as AI-generated by three physicians. Source data are provided as a Source data file.

conformality (lung cohort), or OARs sparing (cervical cohort) than MNL plans. This demonstrates the holistic nature of human review, which considers dose distributions across slices rather than focusing solely on isolated dosimetric metrics (Supplementary Fig.3).

### Generalizability across institutions

To evaluate the generalizability of ATP, the observed inter-institutional variations in assessment results must be understood in terms of patient cohort characteristics and differing evaluation protocols. First, we found that varying fractionation schedules across institutions had no impact on final results, as target doses were derived from prescriptions rather than the models. Second, treatment volumes were generally consistent across the three institutions, with one notable exception: NPC cohorts at Institution C had only half the volume of those at the other institutions (Table 1). This volume difference at Institution C likely correlated with greater deviations in DVH parameters from MNL plans compared to the other two centers, as shown in Fig. 4. The explanation lies in the expected differences in predicted dose falloff patterns between larger and smaller PTVs under similar adherence to OARs sparing guidelines. In Fig. 4, the improved parotid protection at Institution C could stem from the inherent feature of steep dose gradients near parotids in the NPC training sets of Institution A. The inter-institutional bias arising from variations in target definition (to be specific, spatial relationship between OARs and PTV)

https://doi.org/10.1038/s41467-025-67581-z

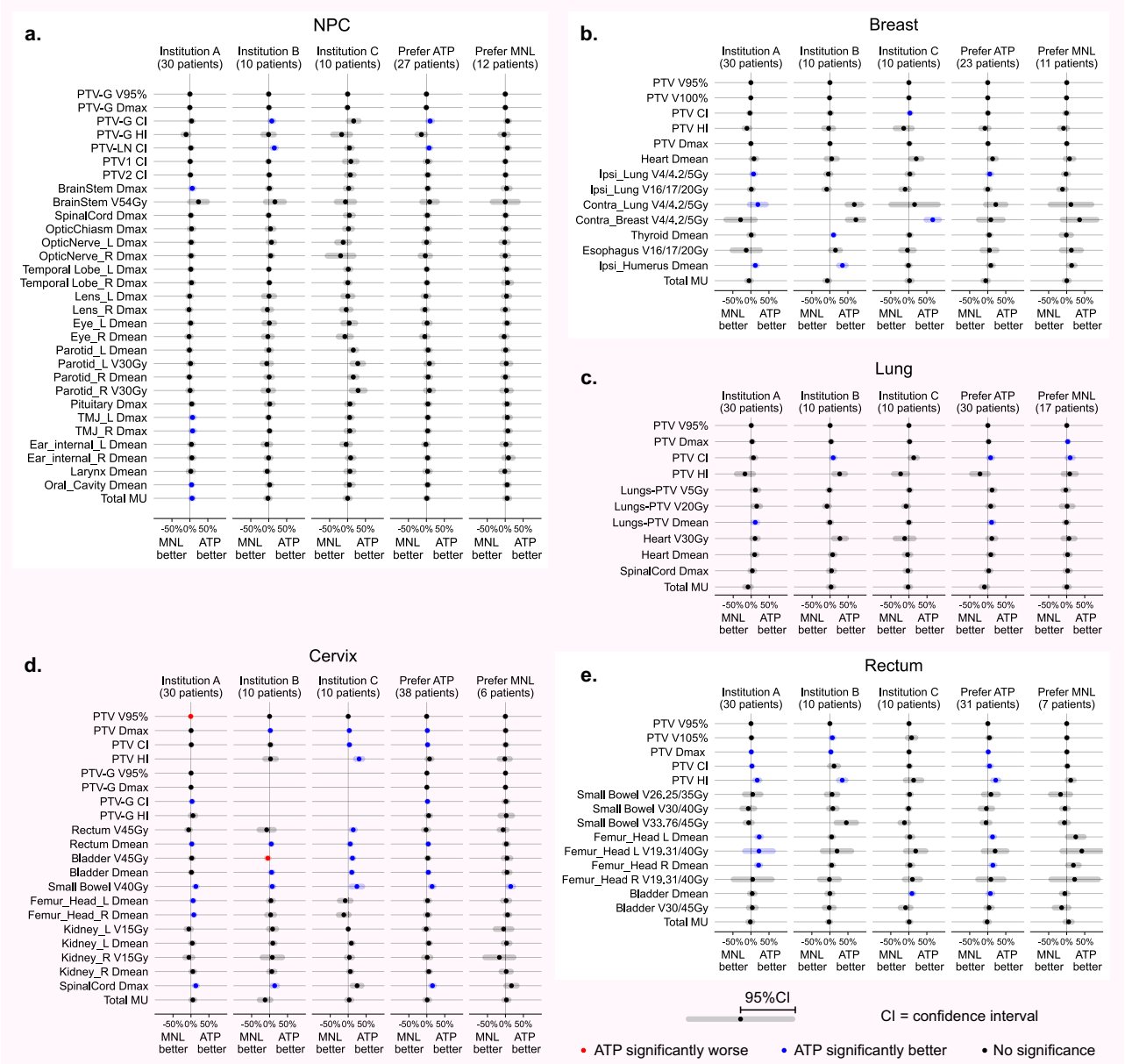


**Fig. 4 | Dot plots of plan quality comparison between ATP and manual planning across five disease sites.** For each site, $n = 30, 10, 10$ cases at Institution A, B, C, respectively. The dots and shaded bars in **a**–**e** represent the average and 95% CI (confidence interval) bounds of the metric difference between the ATP and MNL plans in each institution or in each preferable group (by unanimous selection). Significant differences ($p < 0.05$) are shown in red (ATP worse) or blue (ATP better), while data points of no significance ($p \geq 0.05$) are denoted in black. The $p$ values have been corrected to account for multiple comparisons across numerous dosimetric endpoints. Source data are provided as a Source data file.

can be effectively managed by adjusting clinical goal prediction tolerances. For example, when treatment volumes of NPC site in external institutions quantitatively exceed those in Institution A, predicted parotid doses will unsurprisingly be higher, and comparable sparing can be achieved by simply reducing prediction tolerance (compressing the isodoses near parotids). This measure applies to other tumor sites as well.

Beam geometry variations, namely delivery technique and beam angle configuration, also played a role in multi-institutional performance of ATP. While volumetric-modulated arc therapy (VMAT) was exclusively used in Institution B and C, Institution A primarily employed IMRT (with some exceptions for breast and rectal cases in testing cohort and training datasets; see Table 1 and Supplementary Table 1). Beam setup practices varied not only across institutions but also among individual planners. Since the ATP plans relied on pre-existing MNL beam configurations, using configurations different from the training data could potentially affect the achievable final plans, even with identical predictions.

Similar to comparisons with MNL plans, we compared ATP outputs with predicted dose distributions for the 250 testing cases to investigate the impact of beam configurations and the effectiveness of optimization algorithm, as shown in Fig. 5 and Supplementary Fig.4. The mean voxel-wise errors within body structure exhibited a range of 2.5% to 6.2% across disease types and institutions (Supplementary Fig.4), indicating that ATP plans largely matched dose predictions through effective optimization, despite varying beam configurations. Discrepancies on DVH endpoints of some OARs were observed in Fig.5, with trends roughly consistent among institutions (except for cervix and rectum in C, which used additional isodose controls; see Methods section and

https://doi.org/10.1038/s41467-025-67581-z

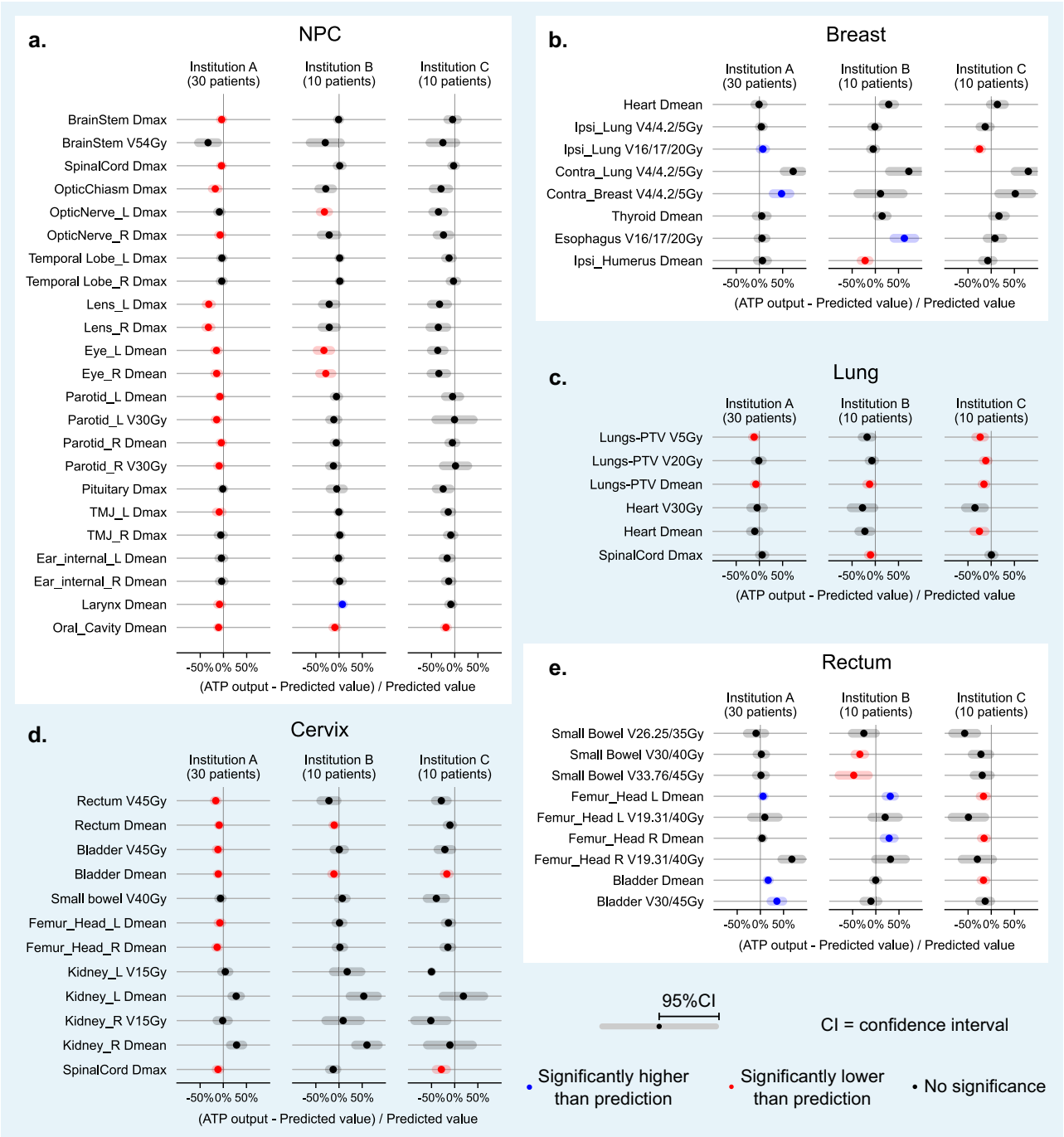


**Fig. 5 | Dot plots of DVH endpoint comparison between the ATP outputs and predicted doses across different disease sites and institutions.** For each site, $n = 30, 10, 10$ cases at Institution A, B, C, respectively. The dots and shaded bars in **a**–**e** represent the average and 95% CI bounds of the metric difference between the final plans and predicted doses in each institution. Significant differences ($p < 0.05$) are shown in red (lower than prediction) or blue (higher than prediction), while data points of no significance ($p \geq 0.05$) are denoted in black. The $p$ values have been corrected to account for multiple comparisons across numerous dosimetric endpoints. Source data are provided as a Source data file.

Supplementary Table 3 for details). This also suggests that beam geometry differences between training and testing populations generally had less impact on final plans than anticipated. The DVH prediction discrepancies were primarily subject to institutional variations in prediction tolerance and priority adjustments, as predicted doses were reshaped during clinical-goal-guided optimization to accommodate tailored goals with different priority levels and tolerances.

To summarize, variations in fractionation schedules, delineation guidelines, beam geometry setups, and evaluation protocols basically had minimal effects on the robustness of final plan generation, emphasizing ATP's generalizability across institutions.

## Further improvement of unfavorable auto-plans

We identified several inadequacies in the less favorable auto-plans, including suboptimal OARs sparing and less conformal dose profiles. These shortcomings stemmed, either directly or indirectly, from inappropriate dose predictions. This indicates that neither the DL models nor fixed goal lists can universally adapt to all patient populations. To address this limitation, various adjustment strategies were

explored within the current framework to better align with clinical judgment.

Thoracic cases from external institutions exemplify the need of such adjustments. Unlike Institution A, which employed a moderate sparing strategy, Institutions B and C prioritized maximal lung protection. For six breast cases rejected by two or more reviewers at Institution B, goal tolerances were refined on a case-by-case basis, achieving plans comparable or superior to MNL plans. These refinements involved a 20–30% reduction in ipsilateral lung V5Gy and V20Gy, balanced by a 10–30% relaxation in tolerances for the mean dose of heart and ipsilateral humeral head, as well as V5Gy for contralateral breast and lung. As shown in Fig. 6a, these adjustments brought the output auto-plans closer to MNL plans across most metrics, while maintaining HI and total MU within clinical guidelines. These findings underscore the feasibility of adjusting goal tolerances to enhance plan quality, and also suggest that institution-specific tailoring of goal lists, as implemented in Institution C, would substantially improve the clinical acceptability and preference for auto-plans.

At Institution C, while OARs sparing in ATP plans was adjusted to levels comparable to that of MNL plans by tailoring goal lists, inferior isodose conformality led to unfavorable clinical judgements, as illustrated in Fig. 6b. This defect was highly case-specific and could not be adequately captured by a single CI parameter. Fine-tuning of the isodose profile was achieved by per-patiently editing the conflicting goals, or by incorporating handcrafted patch contours into the goal list.

As a summary of this subsection, adjusting tolerances of predicted goals is an effective strategy for adapting single-institutional models to meet diverse clinical needs. Our findings demonstrate that tailored goal list at the institutional level significantly enhances the overall acceptability of ATP plans in external validations by accounting for discrepancies in local guidelines. Furthermore, patient-level refinements are helpful for addressing individualized preferences and further improving plan quality.

### Real-world time saving analysis

A typical end-to-end process of radiotherapy treatment planning from plan preparation to final plan production is shown in Fig. 7. Within this framework, the detailed procedures and elapsed time in ATP and MNL planning were compared step-by-step across the five disease types and participating institutions, with the analysis based on a supplementary study involving real-world planning for 75 new patients (see Methods section for study details). While manual interventions remained necessary for plan creation in the present ATP framework, the most time-intensive processes of plan objective specification and iterative tuning were automated and expedited in ATP. This yielded an overall ATP planning time of 10-20 minutes (including necessary per-patient fine-tunings), achieving a 40% reduction in median time compared to MNL-based approaches.

Furthermore, as shown in Fig. 8, the case-by-case analyses of ATP optimization rounds and total time savings indicate that, with predefined initial clinical goals, external institutions generally required more rounds of optimization than internal Institution A to produce acceptable plans in real-world settings. Notably, Institution B needed additional time for iterative plan refinements when using the untailored goal list as a starting point—with some ATP plans even taking longer than MNL plans—whereas Institution C achieved time savings comparable to those of Institution A after moderate tuning rounds. Among the five tumor sites, ATP delivered the greatest time savings for NPC cases, while demonstrating similar time savings for the remaining four sites.

It is noticed that only 64%, 20%, and 28% of ATP plans in Institution A, B, and C, respectively, were deemed acceptable by participating physicists on the first attempt—far lower than the physicians' acceptance rates observed in the retrospective assessments. This discrepancy arises because, beyond adhering to physicians' criteria for plan acceptability, physicists were also subject to additional inspection of plan quality; to ensure first-pass approval, they tended to prioritize developing a plan they perceive as "optimal" through repeated fine-tunings when time allowed. This again underscores the need for flexibility in ATP implementation. Even so, the overall trend of the real-world data remains consistent with the findings of the earlier retrospective study.

## Discussion

The increasing availability of commercial ATP solutions has led to growing interest in their clinical assessment. Recent studies[5–8] show that clinical acceptability rates of ATP within a single optimization reach between 80% and 100% across various treatment sites, with auto-plans outperforming manual plans in 50% to 80% of cases based on retrospective selection. In this study, we introduce an ATP alternative, offering several advantages over previous researches.

The proposed ATP solution provides a coarse-to-fine approach to the perfection of automated planning across diverse clinical settings. To start with, the single-institutional knowledge-based model predicts a generally optimal dose distribution for a new patient anatomy. Subsequently, individualized trade-offs are accounted for by incorporating dynamic adjustments of clinical goals into the rigid dose prediction. In contrast to laborious trial-and-error tuning in manual planning, the present ATP solution directly forwards the optimal goals down into automated optimization to generate a desired final plan ideally in one go, significantly reducing the dedicated time by approximately 40% in real-world settings.

Our multi-institutional assessments demonstrate a general acceptability and a promising preference for the proposed ATP solution across a range of clinical scenarios, aligning with previous studies[5,6]. Notably, the first-pass approval rate in external validations matched that of local validations based on the same dose prediction models, despite the differences in delineation guidelines, beam setup practices, fractionation schedules, and evaluation protocols among the institutions. This achievement underscores the practicality of employing site-specific goal lists that accommodates institutional variability in clinical priorities without necessitating retraining of the models. Furthermore, our approach highlights the effectiveness of modifying inferior plans within the ATP framework to align with clinical preferences, enabling rapid plan generation for both initial treatment and plan adaptation. Compared with other existing ATP methods[5], our solution demonstrates inherent superiority in generalization capacity and flexibility, which are of practical implication for clinical implementation in complex prospective scenarios. In the following discussion, we will further expound on this topic in detail.

As highlighted in previous works[6,21] and confirmed by our findings, ATP has the potential to enhance efficiency, consistency, and standardization in radiotherapy planning while reducing reliance on operator expertise. However, achieving fully automated integration into prospective clinical environments remains challenging[12–14]. This brings us back to the fundamental question posed at the beginning of this article: how can ATP best serve clinical practice? The answer extends beyond plan performance (acceptability) and workflow integration (applicability); it ultimately hinges on the end user[8,22,23].

Drawing insights from automation in aviation[24], the radiation therapy community recognizes the necessity of human oversight in automated systems to navigate subjective clinical judgments and complex trade-offs, including patient-specific factors and quality-of-life considerations[25]. Physician review remains the standard for final decision-making in patient care. ATP's success rate without human involvement may not significantly improve due to the personalized understanding of "clinical optimality"[23]. In head-to-head comparison with manual planning, the perception and trust of treating physicians towards ATP may influence their final decisions for prospective treatment more than the plan performance itself[8].

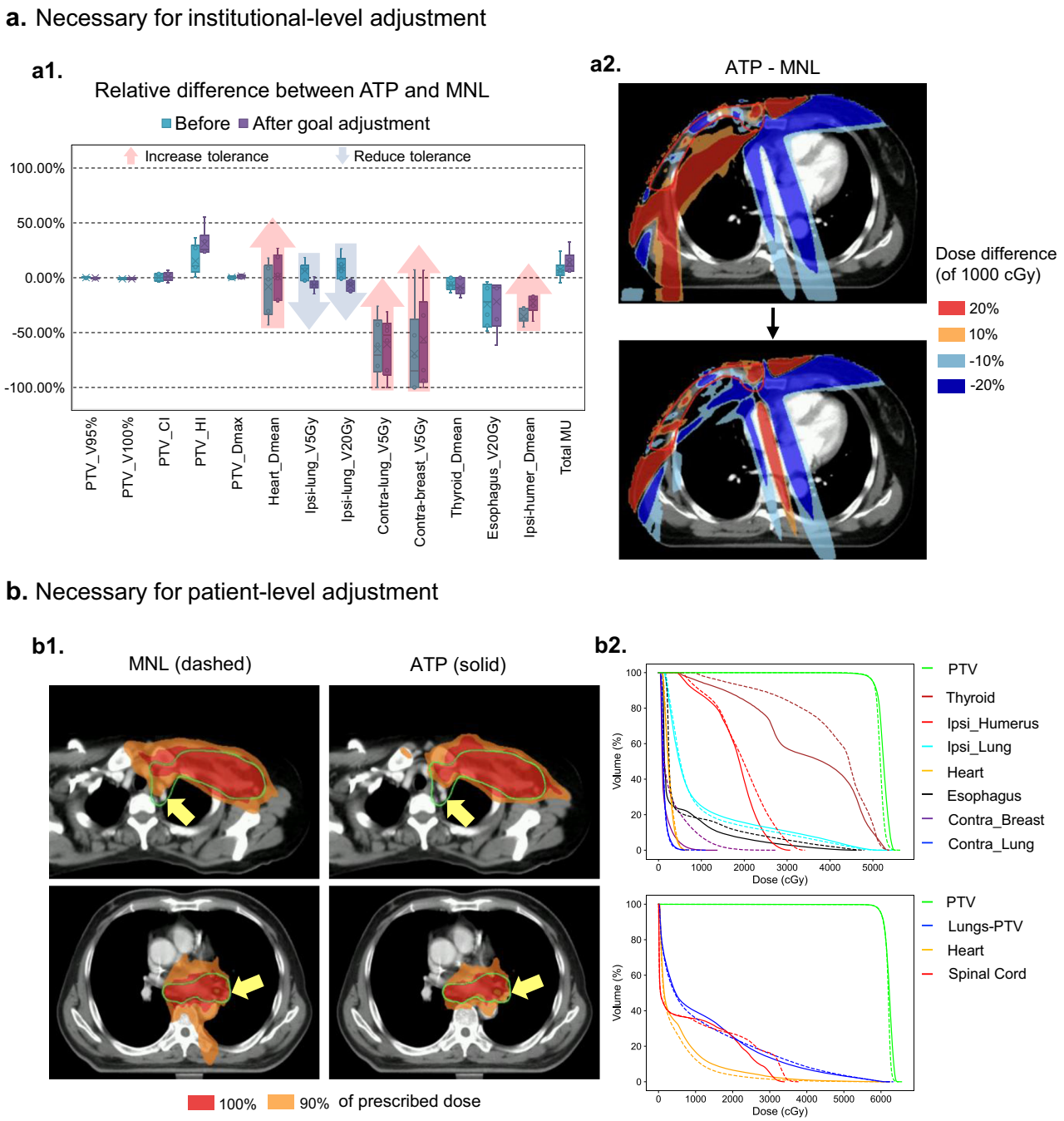


**Fig. 6 | Representative examples of different strategies for further adjustments of unfavorable ATP plans. a** Example presentations necessary for institutional-level adjustments at Institution B: **a1** shows the difference of dosimetric parameters between ATP and MNL plans before and after targeted adjustment of goal list for the inferior breast cases at Institution B ($n = 6$), with data characterized by minima, maxima, median (central line), interquartile range (IQR, box bounds), and whiskers; **a2** shows the dose difference levels before and after adjustments on a representative axial CT image. **b** Example presentations necessary for patient-specific adjustments at Institution C: **b1** displays inferior isodose profiles of the ATP plans (vs. MNL plans) at certain slices for a breast case and a lung case, and **b2** shows the comparable dosimetric metrics of ATP and MNL plans for the two cases. Source data of **a1** are provided as a Source data file.

From the perspective of medical physicists or dosimetrists[26], reverting to manual planning is impractical when ATP output fails to meet clinical needs but is challenging to modify. Rather than a "one-size-fits-all" approach, a flexible and user-friendly manual interface should be integrated into the ATP framework to facilitate seamless plan adjustments[24].

Moreover, ATP's end users face technical and resource-related barriers[26]. Developing models requires extensive training data and testing efforts, which can strain even large institutions and limit accessibility for smaller facilities[27]. Collaborative efforts are essential to adapt models to diverse clinical contexts and ensure broad applicability[28].

The current study offers an effective solution to address these challenges. We further propose a practical integration of single-institutional DL models into a multi-institutional ATP framework, as shown in Fig. 9. This framework involves: (1) Model development: developing DL models and validating recommended goal lists with single-institutional datasets; (2) Retrospective evaluation: assessing

https://doi.org/10.1038/s41467-025-67581-z

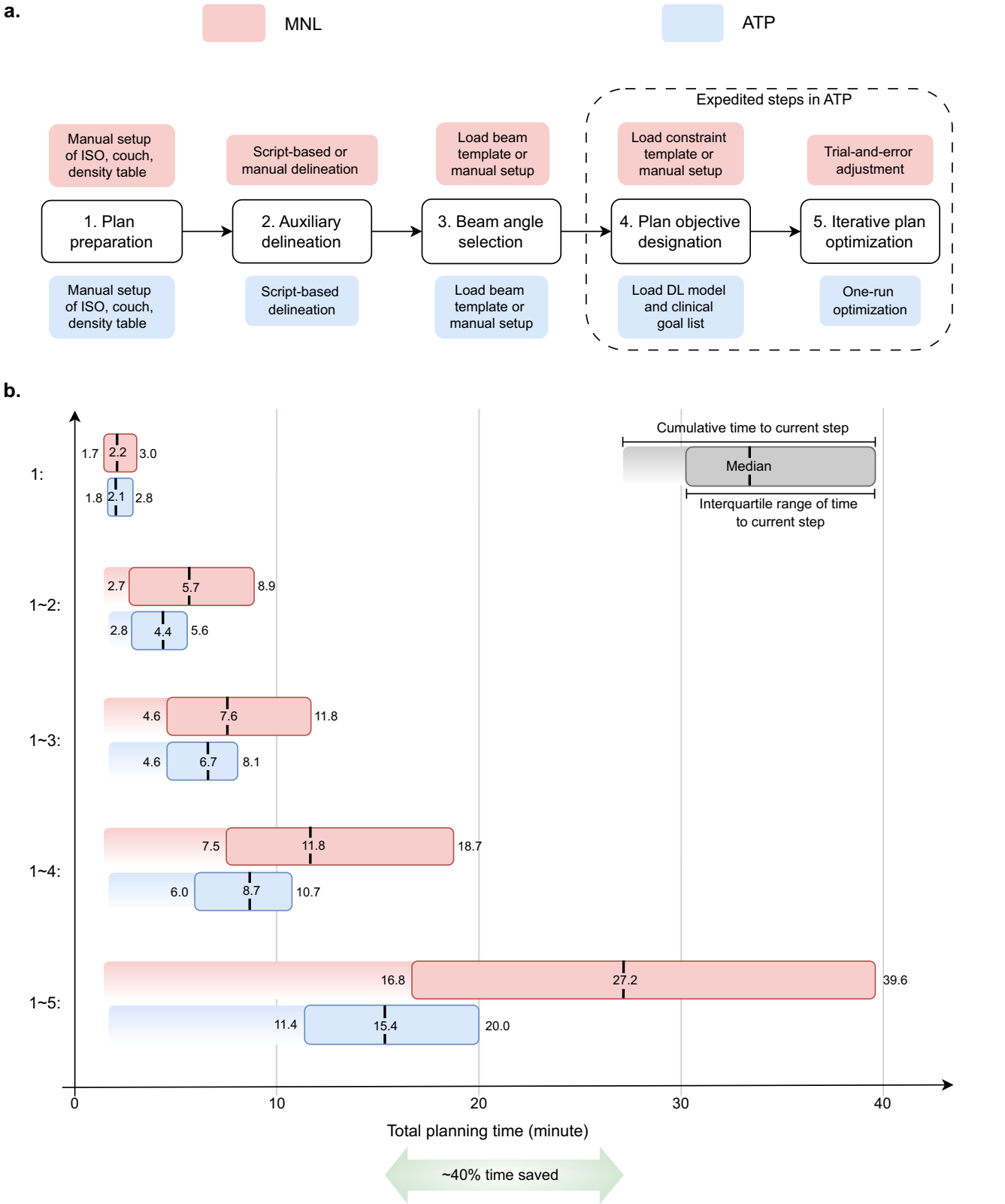


**Fig. 7 | Step-by-step comparison between ATP and MNL planning from plan preparation to final plan production. a** describes the detailed planning procedures and **b** shows the elapsed time to each step. $n = 75$ (5 cases per site at each institution). Source data are provided as a Source data file.

ATP performance at external institutions retrospectively to tailor goal lists to local guidelines and enhance approval rates; (3) Parallel deployment: implementing ATP and manual planning in parallel for blinded physician evaluations to incorporate subjective preferences; and (4) Prospective application: treating patients prospectively with refined auto-plans, incorporating ongoing adjustments as needed. Prior to prospective use, it is essential to establish a threshold for ATP acceptability or preference (e.g.,>80%) in order to proceed to the next phase. For smaller institutions with limited resources of training datasets and computing power, the present ATP framework provides a

https://doi.org/10.1038/s41467-025-67581-z

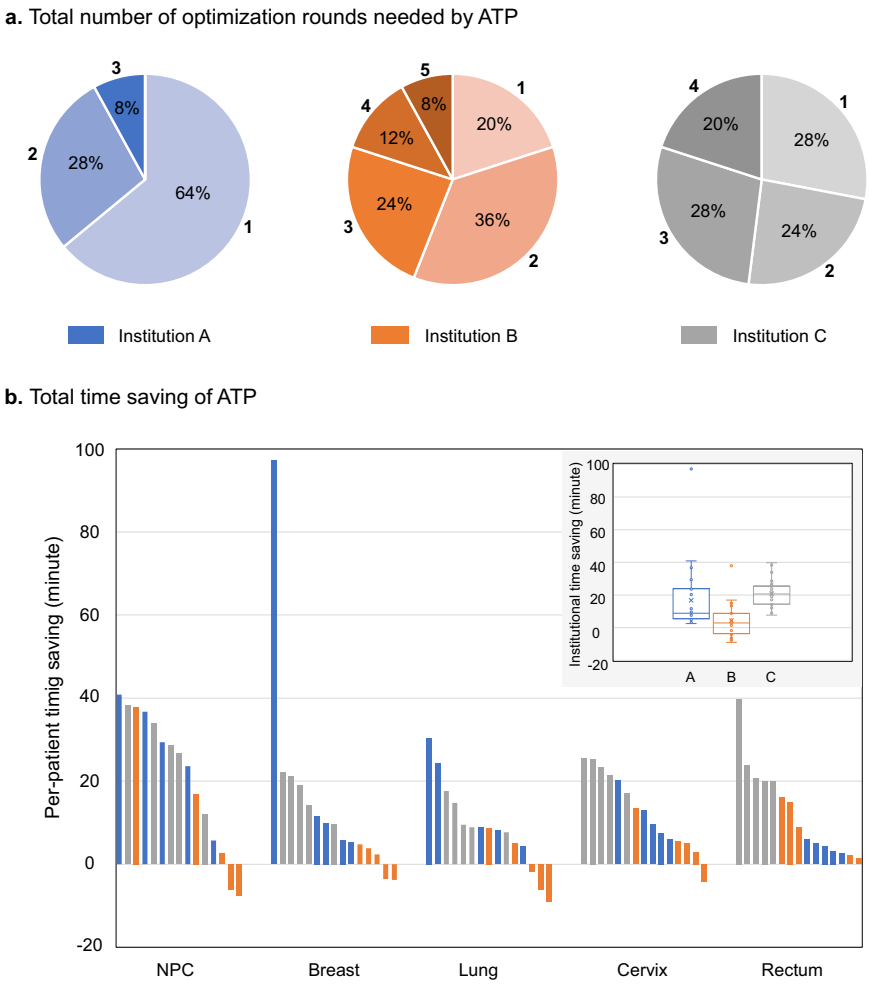


**Fig. 8 | Real-world results of ATP optimization rounds and total time savings on a case-by-case basis.** $n = 75$ (5 cases per site at each institution) **a** Pie chart illustrating the total number of optimization rounds needed by ATP across institutions. **b** Sorted bar chart of total time saving per patient (descending order) across various tumor sites, with a subplot of box chart depicting institutional time savings (data presented as minima, maxima, median, box bounds of IQR, and whiskers). Source data are provided as a Source data file.

feasible approach to leverage the commissioned DL models developed by large centers, namely adopting customized goal lists to incorporate local protocols by following the above steps (2) to (4) in Fig. 9. Besides, local validations with different resource settings will further broaden its clinical contexts and feed back into the development of the ATP solution in a larger scale.

Despite its strengths, our study does face several limitations. The primary challenge is the potential bias introduced by limited patient populations. This bias lies in the knowledge-based nature of the models, which are inherently constrained by the data they were trained on. This limitation is particularly concerning in our study, where models trained on single-institutional data were applied in a multi-institutional context. The key to generalizing the models in the ATP approach is the input of customized clinical goals. Although it provides flexibility in diverse settings, the clinical goal list is intended to establish a template for localized, standardized modifications to accommodate known preferences, e.g., applying fixed prediction tolerance to address differences in target delineation style and evaluation protocol, as elaborated earlier. However, because of their black-box nature, the prediction models also suffer from ambiguous biases that cannot easily be resolved by constant tolerances, necessitating case-specific adjustments and trial-and-error tuning (Fig. 6). This fundamentally hinders improvements in efficiency and consistency.

While generalization capacity has been preliminarily demonstrated, the notably low acceptability of ATP in thoracic sites still indicates substantial room for improvement. The challenge of balancing target coverage with sparing of multiple parallel OARs renders the approach highly sensitive to minor predicted dose fluctuations, even in low-dose regions. Such fluctuations may stem from disparities in tumor locations and beam configurations between training and testing cohorts. Although tumor location diversity was considered during model development, prediction performance remained less than optimal in certain instances (e.g., over-sparing of contralateral breast and ipsilateral humeral head in breast site, Fig. 4), likely attributable to the limited number of training cases per data type. For complex lung cases with highly flexible beam setups, variations in beam configuration can substantially disrupt the delicate trade-offs, leading to suboptimal output.

To mitigate the impacts of population bias, iterative model development utilizing larger, more finely-sorted datasets is essential for adapting to evolving clinical demands. We plan to enhance the thoracic models by incorporating beam configuration as an input variable and developing sub-models labelled with tumor location. Besides, the reliance on manual setting of beam angles impedes full automation of the planning workflow. Future prospective applications will consider integrating pre-configured beam templates or employing DL-based beam orientation optimization to explore potentially

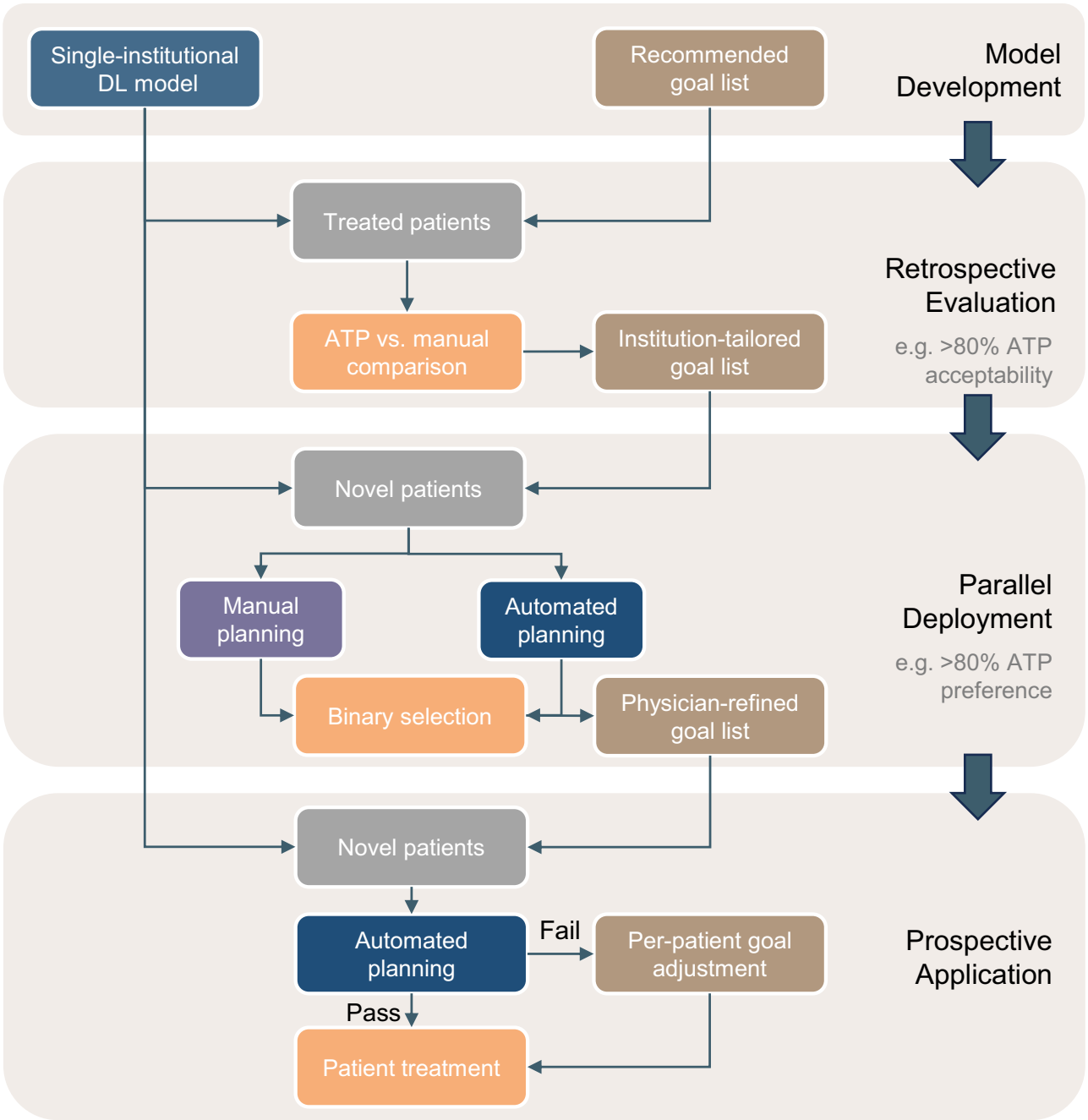


**Fig. 9 | Schematic flowchart for multi-institutional deployment of single-institutional DL models in the proposed ATP framework.** It consists of model development, retrospective evaluation, parallel deployment, and prospective application, where clinical goal list is refined step by step to accomodate local protocols.

superior beam arrangements. Furthermore, an upgraded goal list will facilitate the fine-tuning of the conformality of specific isodose contours by using a distance parameter to the target or organs, thereby minimizing auxiliary contouring during plan adjustment.

Another limitation of this study is the absence of prospective validation of the proposed ATP solution. While retrospective evaluations have confirmed its multi-institutional feasibility, and substantial efficiency improvements have been demonstrated in real-world scenarios, integrating ATP into prospective clinical environments will inevitably encounter additional challenges that require further exploration. Nonetheless, our work provides essential groundwork and practical methodology for future prospective deployment by systematically analyzing potential barriers and solutions during clinical integration. Ongoing development of DL models for additional tumor locations (esophagus, prostate, etc.) aims to expand application scenarios in prospective use. More external validations will be incorporated to investigate robustness against changing circumstances during continuous use, including novel dose optimization algorithms, dose model migration, and hardware variations.

Future research will build on these findings through large-scale prospective multi-disease validations in collaboration with multiple clinical centers. Our goals are to bridge existing gaps and translate our current research into tangible clinical benefits, particularly improved patient outcomes from more consistent treatment plans. As a step forward, the proposed ATP solution has been successfully implemented in an All-in-One radiotherapy workflow for prospective online initial treatment of rectal cancer patients with automated delineation[29] at a single institution[30], and is currently being evaluated in broader

settings. Additionally, integrating an AI agent for ATP generation and modification using large-language models (LLMs) within our framework offers a promising avenue to enhance efficiency and consistency in prospective implementation.

In conclusion, the current study has advanced the automation of radiotherapy treatment planning into real-world clinical practice. The multi-institutional retrospective assessments demonstrate that the proposed ATP solution based on a single-institutional DL framework is noninferior to human-driven planning across various disease sites. Our findings highlight the potential of ATP to facilitate the widespread integration of fully automated radiotherapy treatment planning while addressing its current limitations through iterative improvements and collaborative innovation. The clinically accessible and deliverable ATP solution is expected to enhance efficiency and promote more homogenous treatment plans across regions with varying medical resources. Beyond that, it provides a solid foundation for enabling online replanning in personalized adaptive radiotherapy, paving the way for more efficient and patient-centered treatment workflows.

## Method

### ATP framework in this study

This study was approved by the institutional review board of Institution A (SCCIRB NO. 2201250-16). For retrospective analysis, written informed consents were obtained from all enrolled patients across institutions. The proposed ATP method was jointly developed by our department and the UIH (United Imaging Healthcare Shanghai, China), and has been implemented in its treatment planning system (uTPS, Version R001.3) for clinical use. The ATP workflow consists of two phases, as shown in Supplementary Fig.1. First, in the dose-prediction phase, a voxel-level dose distribution is predicted by a knowledge-based DL model utilizing CT images and contours as input. The model extracts the geometric features of regions of interest (ROI) contours of a specific patient and outputs the dose distribution close to the Pareto optimal plan of the patient, by learning from high-quality historical planning cases. Second, in the auto-optimization phase, DVH indices are extracted from the predicted dose distribution according to a clinical goal list set beforehand, and employed as dose objectives in an inverse optimization. A default site-specific dose control strategy is also incorporated into the optimization process to generate an executable plan. This two-phase workflow enables the ATP solution to efficiently produce high-quality treatment plans. Further details of the ATP solution are provided in subsequent subsections.

**Dose prediction model architecture**. A channel attention densely-connected U-Net (CAD-UNet)[15] was employed in this study to predict the three-dimensional dose distribution of a patient. In this model, the critical ROIs, including PTVs and interested OARs, are treated as the input "channels" for dose prediction. Each individual OAR is assigned to a distinct channel, whereas all PTV regions are aggregated into a single ROI and incorporated into one channel. Regarding the OARs that overlap with PTV, only the doses of the voxels outside the PTV are predicted by the model. The voxels within the PTV are set as the prescribed dose. In cases where a voxel belongs to multiple PTVs with different prescribed doses, the highest prescribed dose is selected as the voxel value for dose prediction. The input contours are converted into matrix-based Boolean masks with a voxel resolution of 3 × 3 × 3 mm, and used as the input of the CAD-UNet with a patch size of $N \times 128 \times 128 \times 128$, where $N$ represents the total number of ROI channels considered for dose prediction.

The CAD-UNet model adopts a U-shape architecture consisting of a contracting path and an expanding path, as illustrated in Supplementary Fig.2. During the contracting process, the channel number of the feature map is doubled by each CAD block. The corresponding feature maps from the contracting path are concatenated with the feature maps in the expanding path, and the number of channels of the feature maps is reduced by CAD blocks. In addition, the shape of the feature maps is doubled through nearest neighbor up-sampling. The model utilizes $1 \times 1 \times 1$ convolution to output the single-channel predicted dose distribution, and used ReLU as the activation function to guarantee the prediction value greater than or equal to 0. To balance the impacts of ROIs with different volumes, a weighted-mean square error (WMSE) loss was applied by introducing volume normalization as follows:

$$L_{WMSE} = \frac{1}{V} \times \sum_{v=1}^{V} w_v \left(\hat{d}^v - d^v\right)^2 \quad (1)$$

where $\hat{d}$ and $d$ represent the predicted dose matrix and ground truth dose matrix. $V$ denotes the total voxel number of output data, and $w$ denotes the weight map of this function. The weight for each voxel in a ROI is the reciprocal of the ROI volume as a percentage of the total voxel number.

**Inverse optimization strategies**. Automatic inverse optimization is performed based on a customized clinical goal list and default dose control strategy for each treatment site. The site-specific goal list comprises of a series of prioritized dosimetric requirements representing the clinical trade-offs between PTV coverage and sparing of different OARs. Structures are differentiated by their planning roles (GTV/CTV/PTV, Organ, Control, etc.). Priority equal to 1 (P1) means the highest priority. Target-related goals are set in priority 2 as default. Priority rules of organ-related goals are described in Supplementary Table 2. For OARs with same or lower priorities than target, the objective values of the DVH indices are extracted from the predicted dose distribution, while hard constraints in P1 require a comprise of PTV coverage and therefore their objective values are specified by user. Control-related goals for dose shaping follow the same rule of priority, with their objective values defined by user, rather than extracted from prediction.

For each predicted goal, a prediction tolerance parameter can be manually set to adjust the objective value in order to generalize to different patient groups and different clinical preferences. For instance, the parameter -t% indicates that the extracted predicted value will be reduced by t% and used as objective. The final clinical goals are mapped to optimization objectives $F_{obj}^{goallist}$, whose weights are determined by the priorities. A robust goal list was obtained by local institution validation on novel patient cohorts (approximately 10 patients per site) until all the auto-plans were comparable to or better than the original clinical plans. Full versions of the recommended goal lists across five sites in Institution A are available in Supplementary Table 3.

Additional dose control strategy is automatically incorporated into the iterative optimization process, including site-specific control of conformality and dose falloff outside the target $F_{obj}^{CI}$, as well as global maximum dose control $F_{obj}^{maxdose}$ (106% of prescribed dose in default). After normalization and summation of the objectives, the total objective function $F_{obj}^{total}$ is presented as follows:

$$F_{obj}^{total} = \sum_{j \in G} F_{obj}^{goallist} + F_{obj}^{CI} + F_{obj}^{maxdose}, \quad (2)$$

where $G$ denotes the total number of goals in the goal list.

### DL model development

The DL models were developed in Institution A for dose prediction of five disease sites, including NPC, lung, breast, cervix, and rectum. In total, historical datasets of 1030 patients were collected for training of the five models. These patients received static/dynamic IMRT (sIMRT/dIMRT) or VMAT treatments in Institution A between 2020 and 2023. All the historical treatment plans were made or approved by senior

physicists so deemed to be clinically optimal. For each site, the patient datasets (ranging from 100 to 300 cases) were randomly divided into training and validation sets, as listed in Supplementary Table 1. The treatment technique of the collected training data was either pure IMRT or pure VMAT, except for breast, with IMRT to VMAT ratio of 1:1.7. It is important to note that the datasets were not subdivided by tumor locations during the model training. Consequently, the breast datasets for both left- and right-sided groups, as well as those with and without regional nodal irradiation, were fed into a single breast model. Similarly, the lung model covered both borderline and central datasets. This measure aimed to generalize the models across a wide range of clinical scenarios for each treatment site.

Model training was conducted using Pytorch on an NVIDIA RTG3090 GPU, by employing Adam optimizer ($\beta1 = 0.9$, $\beta2 = 0.999$) with an initial learning rate of $3 \times 10^{-4}$. We implemented cosine annealing scheduling where the learning rate decayed to a minimum of $1 \times 10^{-6}$. Weight decay (L2 regularization) of $1 \times 10^{-4}$ was applied to prevent overfitting. Each epoch consisted of 500 iterations with a batch size of 2. Early stopping (patience=50 epochs) monitored the validation loss, with maximum training limited to 1000 epochs. He_normal initialization method[31] was used to initialize the network parameters. Data augmentation was implemented in real-time during training to enhance model generalizability. The augmentation pipeline incorporated three primary geometric transformations: axial translation with a displacement range of ±64 voxels (corresponding to ±19.2 mm given the 3 $mm^3$ voxel resolution), axial scaling applied uniformly across all axes within ±10% of the original dimensions, and three-dimensional rotation with angular variations of ±10° around each spatial axis. These transformations were dynamically generated using fixed random seeds (seed = 2020) to ensure reproducible augmentation patterns across different training sessions, with transformation parameters sampled from uniform distributions at each iteration. The computational efficiency of this approach enabled on-the-fly processing without requiring pre-augmented data storage, while maintaining batch diversity through probabilistic application of combined transformations.

### Patient cohorts

Retrospective comparison of ATP and manual planning was performed in three cancer institutions of China, including the internal institution (A), and two external institutions (B and C). A total of 250 patients were enrolled in this study, with 50 patients per site. Patients with artificial femoral head or hip replacement implant were excluded for cervical and rectal studies, and no patients were excluded for the other studies. Sex or gender was not considered in the study design because sex or gender was not a relevant factor to the evaluation of dosimetric performance and time saving of automated radiotherapy planning. Detailed characteristics of the patient cohorts are summarized in Table 1. PTVs and OARs necessary for dose prediction were contoured according to the respective delineation guideline of local institution, and the MNL plans were generated and clinically approved beforehand in uTPS for all the enrolled patients.

### ATP plan generation

For retrospective comparison, the ATP plans were generated in the following steps by manual execution: create a new plan by copying the beam configuration of the MNL plan, load the prediction model and clinical goal list of the appropriate site, and start optimization using the ATP module implemented in uTPS. To investigate the robustness and generalizability of the ATP solution, different strategies of clinical goals were adopted in three institutions. The Institution A and B shared the same goal lists, while Institution C modified the goal lists to align with its local clinical standards, as detailed in Supplementary Tables 3 and 4. No per-patient goal adjustments were applied during the plan generation.

In a prospective scenario, manual operations on localization, couch replacement, density table selection, and beam orientation setup are currently needed before ATP plan generation, as shown in Fig. 7. A fully automated scripted pipeline can be available by using pre-configured plan templates[30].

### Plan evaluation and statistical analysis

The ATP and MNL plans were compared in subjective and objective criteria, respectively. For subjective evaluation, three expert physicians from each institution were invited to evaluate the plans of a certain site. Source of the plans (ATP or MNL) was blinded to the reviewing physicians. The physicians were required to fill a questionnaire of a 3-point Likert scale (agree, disagree, or cannot judge), regarding the clinical acceptability, preferability, and generation method of the plans. Selection of "cannot judge" indicates equivalence or indistinguishability, so it scores as ATP preference if the plan is deemed acceptable, and as misidentification of ATP generation. Physicians were also prompted to provide comments or reasons for their selections. We summarized the keywords mentioned in the comments of the reviewing physicians and plotted a snapshot of word cloud in Supplementary Fig.3.

Dosimetric comparison between the ATP group and MNL group was performed at the DVH endpoints defined by the RTOG protocols of each site[16–20]. Difference in dose fractionation was accounted for by converting the evaluating isodoses to the corresponding equivalent doses in 2 Gy/fraction (EQD2). The results were extracted in bulk using a Python-based in-house tool (RadDOP Version 2.5.0), and statistically analyzed with a two-tailed paired-sample $t$ test for significance ($p < 0.05$) using Microsoft Office Excel (Version 16.93.1). According to the results of blinded review collected from all the institutions, the dosimetric parameters were also compared between the ATP-preferred group and MNL-preferred group. The relative difference of each DVH metric was found by $(D_{MNL} - D_{ATP})/D_{MNL} \times 100\%$ (V95% and CI of PTV had opposite sign to show ATP improvement). The average value of the differences and the 95% CI were calculated for each metric and shown in Fig. 4. To account for multiple comparisons across numerous dosimetric endpoints, the initial computed $p$ values ($p_i$) were adjusted by applying Bonferroni correction, namely for simultaneous testing with $m$ dosimetric endpoints, the adjusted $p$ value of each comparison $p_i'$ was written as,

$$p_i' = min\{p_i \times m, 1\}\ (1 \le i \le m). \tag{3}$$

### ATP output versus dose prediction

The ATP outputs were also compared with the predicted dose distributions to investigate the robustness and effectiveness of the automated optimization algorithm in various clinical settings. Differences in both voxel wise (within body structure) and DVH endpoints (for OARs only) were extracted for the 250 testing cases, as shown in Supplementary Fig. 4 and Fig. 5, respectively. The voxel-wise error shown was calculated as the absolute dose difference relative to the prescription dose, i.e., $|D_{ATP} - D_{predit}|/D_{prescription} \times 100\%$, while the comparison of DVH endpoints was evaluated by $(D_{ATP} - D_{predict})/D_{predict} \times 100\%$.

The observed discrepancies in DVH endpoints between the predicted and final plans can be attributed to the following factors. First of all, these discrepancies primarily stemmed from the clinical-goal-guided optimization with different prediction tolerances and priorities. For example, the clinical goals of maximum dose limits in priority 1 for serial OARs in NPC, as well as additional isodose controls in priority 1 for cervical and rectal cases at Institution C (Supplementary Table 3), led to significantly lower doses of OARs than prediction. Particularly, compared to A and B, a different trend of DVH deviations in Institution C was subject to the tolerance adjustments of tailored goals, i.e., reduced lung sparing in breast and lung sites

(Supplementary Table 4). Difference in delivery techniques across institutions (pure IMRT in A versus pure VMAT in B and C for NPC, lung, cervix) may contribute to deviations from prediction in low-dose regions, but generally seems to have minimal effect on the DVH parameters of final plans.

### ATP plan improvement

The OARs sparing and isodose coverage of the unfavorable ATP plans were improved within the current framework by editing the goal lists. Specifically, inferior OARs sparing was addressed by adjusting the tolerance of the predicted goal in both institution level and patient level, as illustrated in Fig. 6a. Conformality of a specific isodose profile was adjusted in patient level only, as shown in Fig. 6b. In the present method, the isodose surface around the target was shaped by both isotropic dose falloff in an intrinsic rate and dose constraints to influencing organs. If there was a need to change the overall dose gradient, additional ring structures had to be generated and incorporated into the goal list (Supplementary Table 3). As to locally modifying the isodose contour at a certain slice, such as cold/hot spot inside the target and dose extruding/notching outside the target, usually it can be solved by redressing the conflicting constraints in the goal list. In extreme cases, a handcrafted patch contour, similar to manual planning, can be delineated and included in the goal list for further optimization.

A cold start of plan re-optimization was usually recommended when predicted objectives were adjusted, in order to obtain a better output by reoptimizing the fluence map. Although a warm start could be used for faster re-optimization (<2 min), it was primarily reserved for patch modifications. Besides, it is worth noting that the intended prediction adjustment does not necessarily result in changes to the final result, as the priority order of the goals has to be primarily observed.

### Timing analysis of end-to-end planning

Herein, we analyzed the timing of the end-to-end planning process to evaluate the efficiency improvement of the current ATP method in contrast to manual approach. Due to the retrospective nature of this study, the exact time spent on the historically manual plans used in the paper was not available. To measure the actual time savings achieved by ATP, we conducted a small-scale supplementary study about real-world planning time comparison by recreating new MNL and ATP plans across the three institutions. The study design is detailed as follows.

A total of 75 patients were included in this supplementary study, with five patients allocated to each tumor site at each institution. The enrolled patients were selected from new patients waiting for clinical treatment between August and September 2025, so that the real-world timelines were recorded. In each institution, the MNL plans were created by five senior physicists (5-year experience or more) to include inter-planner variability (each physicist developed five plans, one for each tumor site). The corresponding ATP plans were generated (with delivery techniques remaining unchanged) by another senior physicist who was familiar with the automated system. The participants were required to follow their own routines of planning (either pure-manual or template/script-based executions) and record the actual time expended in each step described in Fig. 7. All the physicists were blinded to each other's outputs. For a fair comparison, both MNL and ATP plans were fine-tuned, if necessary, until the plans were considered acceptable and ready for physician evaluation, the timepoint of which depended on physicist's judgement. The number of fine-tuning rounds of ATP plans was also recorded (both cold- and warm-start re-optimization were included).

The calculated median and interquartile range of the elapsed time to each step, as well as the cumulative time range of the entire cohort, are presented in Fig. 7. Figure 8 displays the results of ATP optimization rounds and total time savings on a case-by-case basis across various institutions and disease sites.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Due to privacy concerns and regulation requirements, raw patient data for model development and plan evaluation in this study cannot be shared and are not publicly available. The processed data are provided within the Article, Supplementary Information or Source Data file. Also because of policy considerations, requests for DL dose prediction models should be directed to the corresponding author W.H. and made available after specific REB approvals and bespoke data sharing agreement established between the institution and the requesting party. Source data are provided with this paper.

## Code availability

The ATP algorithm is available with the licensed automated treatment planning module in uTPS (United Imaging Healthcare, Shanghai, China). The installation package of RadDOP software used for batch extraction of dosimetric metrics of radiotherapy plans is available on GitHub at https://github.com/HuangShiXiong9146/Radiotherapy-Dosimetry-Omics-Platform-RadDOP-[32]. The version used in this study can be found at Zenodo repertory through https://doi.org/10.5281/zenodo.17490334[32].

## References


1. Nelms, B. E. et al. Variation in external beam treatment plan quality: an inter-institutional study of planners and planning systems. *Pract. Radiat. Oncol.* **2**, 296–305 (2012).
2. Moore, K. L. et al. Quantifying unnecessary normal tissue complication risks due to suboptimal planning: a secondary study of RTOG 0126. *Int. J. Radiat. Oncol. Biol. Phys.* **92**, 228–235 (2015).
3. Nguyen, D. et al. Advances in automated treatment planning. *Semin. Radiat. Oncol.* **32**, 343–350 (2022).
4. Hussein, M., Heijmen, B. J. M., Verellen, D. & Nisbet, A. Automation in intensity modulated radiotherapy treatment planning-A review of recent innovations. *Br. J. Radio.* **91**, 20180270 (2018).
5. Meyer, P. et al. Automation in radiotherapy treatment planning: examples of use in clinical practice and future trends for a complete automated workflow. *Cancer Radiother.* **25**, 617–622 (2021).
6. Cornell, M. et al. Noninferiority study of automated knowledge-based planning versus human-driven optimization across multiple disease sites. *Int J. Radiat. Oncol. Biol. Phys.* **106**, 430–439 (2020).
7. Wheeler, P. A. et al. Multi-institutional evaluation of a Pareto navigation guided automated radiotherapy planning solution for prostate cancer. *Radiat. Oncol.* **19**, 45 (2024).
8. McIntosh, C. et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat. Med.* **27**, 999–1005 (2021).
9. Jiang, C., Ji, T. & Qiao, Q. Application and progress of artificial intelligence in radiation therapy dose prediction. *Clin. Transl. Radiat. Oncol.* **47**, 100792 (2024).
10. Szalkowski, G., Xu, X., Das, S., Yap, P. T. & Lian, J. Automatic treatment planning for radiation therapy: a cross-modality and protocol study. *Adv. Radiat. Oncol.* **9**, 101649 (2024).
11. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
12. Zadnorouzi, M. & Abtahi, S. M. M. Artificial intelligence (AI) applications in improvement of IMRT and VMAT radiotherapy treatment planning processes: a systematic review. *Radiography* **30**, 1530–1535 (2024).

13. Gooding, M. J. et al. Fully automated radiotherapy treatment planning: a scan to plan challenge. *Radiother. Oncol.* **200**, 110513 (2024).
14. Amaloo, C., Hayes, L., Manning, M., Liu, H. & Wiant, D. Can automated treatment plans gain traction in the clinic?. *J. Appl Clin. Med. Phys.* **20**, 29–35 (2019).
15. Liu, R. et al. A new deep-learning-based model for predicting 3D radiotherapy dose distribution in various scenarios[C]//2020. In *Proc. 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics* (CISP-BMEI, 2020).
16. Lee, A. W. et al. International guideline on dose prioritization and acceptance criteria in radiation therapy planning for nasopharyngeal carcinoma. *Int. J. Radiat. Oncol. Biol. Phys.* **105**, 567–580 (2019).
17. Bradley, J. D. et al. Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study. *Lancet Oncol.* **16**, 187–199 (2015).
18. Bazan, J. G. & White, J. R. The role of postmastectomy radiation therapy in patients with breast cancer responding to neoadjuvant chemotherapy. *Semin Radiat. Oncol.* **26**, 51–58 (2016).
19. Pötter, R. et al. The EMBRACE II study: the outcome and prospect of two decades of evolution within the GEC-ESTRO GYN working group and the EMBRACE studies. *Clin. Transl. Radiat. Oncol.* **9**, 48–60 (2018).
20. Hong, T. S. et al. NRG oncology radiation therapy oncology group 0822: a phase 2 study of preoperative chemoradiation therapy using intensity modulated radiation therapy in combination with capecitabine and oxaliplatin for patients with locally advanced rectal cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **93**, 29–36 (2015).
21. Wang, J. et al. Is it possible for knowledge-based planning to improve intensity modulated radiation therapy plan quality for planners with different planning experiences in left-sided breast cancer patients?. *Radiat. Oncol.* **12**, 85 (2017).
22. Lenharo, M. The testing of AI in medicine is a mess. Here's how it should be done. *Nature* **632**, 722–724 (2024).
23. Baroudi, H. et al. Automated contouring and planning in radiation therapy: what is 'clinically acceptable'?. *Diagnostics* **13**, 667 (2023).
24. Callens, D. et al. Is full-automation in radiotherapy treatment planning ready for take off? *Radiother Oncol.* **201**, 110546 (2024).
25. Alborghetti, L. et al. Selective sparing of bladder and rectum subregions in radiotherapy of prostate cancer combining knowledge-based automatic planning and multicriteria optimization. *Phys. Imaging Radiat. Oncol.* **28**, 100488 (2023).
26. Petragallo, R., Bardach, N., Ramirez, E. & Lamb, J. M. Barriers and facilitators to clinical implementation of radiotherapy treatment planning automation: a survey study of medical dosimetrists. *J. Appl. Clin. Med. Phys.* **23**, e13568 (2022).
27. Shah, P. et al. Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ Digit Med.* **2**, 69 (2019).
28. Yang, J. et al. Generalizability assessment of AI models across hospitals in a low-middle and high income country. *Nat. Commun.* **15**, 8270 (2024).
29. Shi, F. et al. Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. *Nat. Commun.* **13**, 6566 (2022).
30. Yu, L. et al. Technical note: first implementation of a one-stop solution of radiotherapy with full-workflow automation based on CT-linac combination. *Med Phys.* **50**, 3117–3126 (2023).
31. He, K., X. Zhang, X., Ren S. & Sun, J. "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," In *Proc. IEEE International Conference on Computer Vision (ICCV)* 1026–1034 (IEEE, 2015).
32. Yu, L. et al. Multicenter study on the versatility and adoption of AI-driven automated radiotherapy planning across cancer types. *Radiotherapy Dosimetry Omics Platform RadDOP* https://doi.org/10.5281/zenodo.17490334 (2025).

## Acknowledgements

This work is supported by the National Key Research and Development Program of China (2024YFC2418400 to Z.Z. and 2022YFC2404603 to W.H.), National Natural Science Foundation of China (12505393 to L.Y., 12475339 to J.W., and 12405380 to Y.Z.), Shanghai Committee of Science and Technology Fund (25TS1405300 to Z.Z.), and the Key Research and Development Program of Hunan Province (2025JK2138 to Q.N).

## Author contributions

Study conception and design: Z.Z., J.W., and W.H.; Data collection and analysis: L.Y., Q.N., B.W., S.H., Y.Z., and Y.G.; Interpretation of results: Q.N., G.S., and J.W.; Manuscript preparation: L.Y., J.W., F.S., K.Z., and W.H. All authors reviewed the results and approved the final version of the manuscript. L. Yu, Q. Ni, and B. Wang contributed equally to this work.

## Competing interests

K.Z. is the employee of Shanghai United Imaging Healthcare Co., Ltd.; F.S. is the employee of Shanghai United Imaging Intelligence Co., Ltd. The two authors mentioned above contributed to the revision of the manuscript. They had no role in the study design, data analysis, or result interpretation of this work. The remaining authors declare no competing interests relevant to this article. All authors state that they have no financial interests to disclose.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-67581-z.

**Correspondence** and requests for materials should be addressed to Zhen Zhang, Jiazhou Wang or Weigang Hu.

**Peer review information** *Nature Communications* thanks Jan Peeken, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.