MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™ Part of the MACHINE LEARNING Series

PAPER • OPEN ACCESS

# Performance deterioration of deep learning models after clinical deployment: a case study with auto-segmentation for definitive prostate cancer radiotherapy

To cite this article: Biling Wang *et al* 2024 *Mach. Learn.: Sci. Technol.* **5** 025077

View the article online for updates and enhancements.

You may also like

- Machine-learning strategies for the accurate and efficient analysis of x-ray spectroscopy
Thomas Penfold, Luke Watson, Clelia Middleton et al.
- Deep learning methods for Hamiltonian parameter estimation and magnetic domain image generation in twisted van der Waals magnets
Woo Seok Lee, Taegeun Song and Kyoung-Min Kim
- Explainable machine learning assisted molecular-level insights for enhanced specific stiffness exploiting the large compositional space of AlCoCrFeNi high entropy alloys
K K Gupta, S Barman, S Dey et al.

This content was downloaded from IP address 173.177.195.252 on 27/02/2026 at 13:11

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 025077 https://doi.org/10.1088/2632-2153/ad580f

MACHINE LEARNING
Science and Technology

CrossMark

OPEN ACCESS

RECEIVED
12 March 2024

REVISED
6 May 2024

ACCEPTED FOR PUBLICATION
13 June 2024

PUBLISHED
25 June 2024

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



PAPER

# Performance deterioration of deep learning models after clinical deployment: a case study with auto-segmentation for definitive prostate cancer radiotherapy

**Biling Wang**[1,3,5], **Michael Dohopolski**[1,2,5], **Ti Bai**[1,2], **Junjie Wu**[1,2], **Raquibul Hannan**[1,2], **Neil Desai**[1,2], **Aurelie Garant**[1,2], **Daniel Yang**[1,2], **Dan Nguyen**[1,2], **Mu-Han Lin**[1,2], **Robert Timmerman**[1,2], **Xinlei Wang**[3,4,*] and **Steve B Jiang**[1,2,*]

[1] Medical Artificial Intelligence and Automation Laboratory, University of Texas Southwestern Medical Center, Dallas, TX, United States of America
[2] Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, United States of America
[3] Department of Statistics and Data Science, Southern Methodist University, Dallas, TX, United States of America
[4] Department of Mathematics, University of Texas at Arlington, Dallas, TX, United States of America
[5] Co-first authors.
[*] Authors to whom any correspondence should be addressed.

**E-mail:** xinlei.wang@uta.edu and steve.jiang@utsouthwestern.edu

**Keywords:** deep learning, segmentation, model performance deterioration, radiotherapy

Supplementary material for this article is available online

## Abstract

Our study aims to explore the long-term performance patterns for deep learning (DL) models deployed in clinic and to investigate their efficacy in relation to evolving clinical practices. We conducted a retrospective study simulating the clinical implementation of our DL model involving 1328 prostate cancer patients treated between January 2006 and August 2022. We trained and validated a U-Net-based auto-segmentation model on data obtained from 2006 to 2011 and tested on data from 2012 to 2022, simulating the model's clinical deployment starting in 2012. We visualized the trends of the model performance using exponentially weighted moving average (EMA) curves. Additionally, we performed Wilcoxon Rank Sum Test and multiple linear regression to investigate Dice similarity coefficient (DSC) variations across distinct periods and the impact of clinical factors, respectively. Initially, from 2012 to 2014, the model showed high performance in segmenting the prostate, rectum, and bladder. Post-2015, a notable decline in EMA DSC was observed for the prostate and rectum, while bladder contours remained stable. Key factors impacting the prostate contour quality included physician contouring styles, using various hydrogel spacers, CT scan slice thickness, MRI-guided contouring, and intravenous (IV) contrast ($p < 0.0001$, $p < 0.0001$, $p = 0.0085$, $p = 0.0012$, $p < 0.0001$, respectively). Rectum contour quality was notably influenced by factors such as slice thickness, physician contouring styles, and the use of various hydrogel spacers. The quality of the bladder contour was primarily affected by IV contrast. The deployed DL model exhibited a substantial decline in performance over time, aligning with the evolving clinical settings.

## 1. Introduction

Over the last ten years, artificial intelligence (AI), driven by deep learning (DL) techniques, has achieved remarkable progress, particularly in areas such as computer vision (CV) and natural language processing (NLP), leading to transformative developments across numerous applications. This surge has led to significant enthusiasm in the medical realm, and DL-related medical publications have been growing exponentially since 2015 [1]. However, despite the promising prospects of DL in the medical field, its practical deployment remains constrained [2].

© 2024 The Author(s). Published by IOP Publishing Ltd

This lack of clinical translation is multifactorial. First, the interpretability of many DL models remains a challenge [3, 4]; therefore, clinicians are rightly skeptical when assessing whether a model can be appropriately applied to patient care [5–9]. This skepticism is not unfounded, as issues of generalizability persist in many DL models [10–20]. For instance, a model trained and validated on data from one institution may fail when implemented at another [21]. Another critical issue, often overlooked in medical literature, is the potential for a model's performance to degrade after initial deployment [22–26]. The decline in model performance post-clinical deployment can often be attributed to data drift, such as variations in imaging acquisition protocols over time within the institution, and evolving practice patterns as new faculty join [27].

In one of the first clinically-oriented studies evaluating a model's performance, Davis *et al* [22] observed a temporal decline in their model's ability to predict acute kidney injury. While they attributed this decline to calibration drift, they did not explore the underlying factors in detail. Similarly, Nestor *et al* [25] also noted temporal performance changes when predicting mortality and prolonged length of stay. Clearly, there is a pressing need for further research to explore how and why a DL model's performance may deteriorate over time.

In this study, we hypothesized that the efficacy of pretrained DL models might diminish over time post-deployment, particularly if the operational context shifts. Given the nascent stage of DL model deployment in real-world settings, our investigation necessitated a retrospective simulation of the clinical deployment process. We showcase the impact of evolving clinical workflows on model performance to stress the need for continuous evaluation even after initial model validation.

## 2. Methods and materials

We retrospectively simulated the clinical implementation of our DL model to investigate temporal performance patterns. Our study involved 1328 prostate cancer patients who underwent definitive external beam radiotherapy (EBRT) between January 2006 and August 2022 at the UT Southwestern Medical Center (UTSW). We trained a U-Net-based auto-segmentation model on data obtained between 2006 and 2011 and tested it on data obtained between 2012 and 2022, simulating the model's clinical deployment starting in 2012. We measured the model's performance using the Dice similarity coefficient (DSC), visualized the trends in contour quality (DSC) using exponentially weighted moving average (EMA) curves. Additionally, we performed Wilcoxon Rank Sum Test to analyze the differences in DSC distributions across distinct periods, and multiple linear regression to investigate the impact of various clinical factors.

### 2.1. Dataset

In this single-institutional study approved by our institutional review board, we identified 1480 patients at UTSW diagnosed with prostate cancer and treated with definitive EBRT from January 2006 to August 2022. EBRT treatment regimens included conventional, moderately hypofractionated, or ultra-fractionated radiotherapy, also known as stereotactic body radiotherapy (SBRT). All patients had delineated contours on radiotherapy planning computed tomography (CT) for the prostate, rectum, or bladder. We excluded prostate contours that incorporated the seminal vesicles. To be included, patients were required to have at least a prostate, bladder, or rectum contour. Moreover, patients were excluded if significant artifact was observed. Our final cohort comprised 1328 patients (figure 1).

Within the final cohort, 982 had well-defined prostate contours, 1269 had available rectum contours, and 1277 had available bladder contours. One hundred and sixty three (163) patients were treated between 2006 and 2011, 203 patients between 2012 and 2014, 602 between 2015 and 2019, and 360 patients were treated between 2020 and 2022.

We extracted variables including treating physicians, slice thicknesses for the CT scans, types of hydrogel spacers, either non-contrast (Type I spacer) or contrast-enhancing (Type II spacer), use of IV contrast at time of CT sim (measured by evaluating IV contrast was present in the bladder), and Magnetic Resonance Imaging guided (MRI-guided) contouring techniques. The detailed information of these variables can be seen in section 3.2.

### 2.2. Model training

We utilized a 3D U-Net-based auto-segmentation model to contour the prostate, rectum, and bladder on CT images intended for radiotherapy planning. Our implementation was based on the open-source MONAI U-Net [28]. During training, we enhanced our dataset by randomly rotating ($\pm$15 degrees), scaling (0.8–1.2 times), and shifting ($\pm$64 pixels) the images in all three dimensions. We set the batch size to one. Then the model was trained using the Adam optimizer with default hyperparameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) over $1 \times 10^5$ iterations, leveraging the dice loss function. We initiated the learning rate at $1 \times 10^{-4}$, reducing it to $1 \times 10^{-5}$ at the $4 \times 10^4$th iteration and $1 \times 10^{-6}$ at the $8 \times 10^4$th iteration, respectively. We randomly

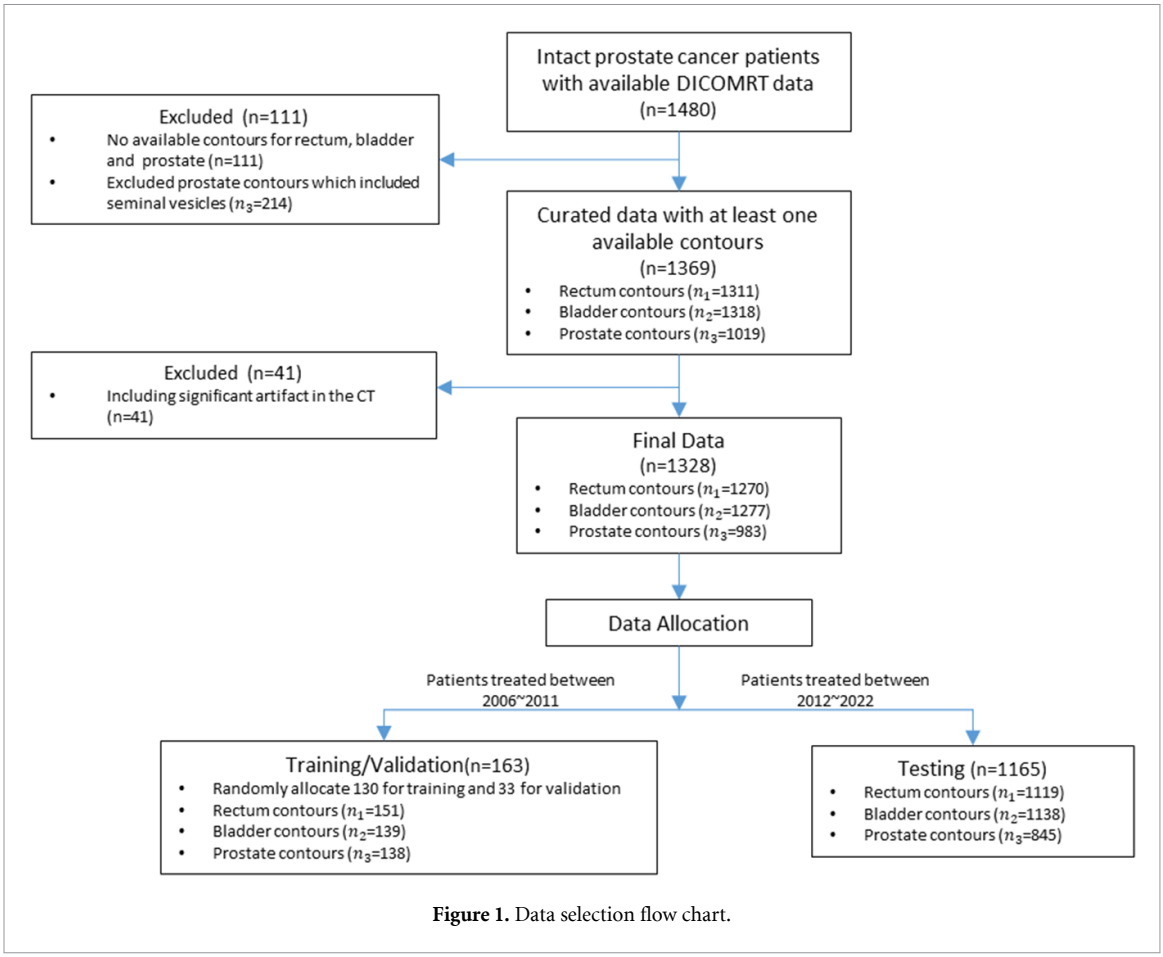



**Figure 1.** Data selection flow chart.

allocated 130 patients for training and 33 patients for validation from the 163 patients treated before 2012. The final model was selected based on the based validation performance.

### 2.3. Longitudinal performance evaluation

DSC was calculated to evaluate the model's performance (i.e. the quality of the contours it generated). The ground truth contours were the clinical contours used in the patients' delivered treatments.

To comprehensively evaluate the model's performance and discern any influencing factors, we employed both visualization techniques and statistical analyses. We charted the model's performance trends from 2012 to 2022 using an EMA. This was applied to the DSC trend curve (termed EMA DSC) with a window size of 180 d and a minimum of 50 observations within the window to have a value. Detailed computational procedures for EMA DSC are provided in the supplementary file. We performed Wilcoxon Rank Sum Test to analyze the differences in DSC distributions across distinct periods.

Our analysis delved into the potential effects of evolving clinical practices on the model's temporal performance. We meticulously examined the influence of factors such as CT imaging slice thickness, types of hydrogel spacers, IV contrast, MRI-guided contouring techniques, and physician styles. We utilized multiple linear regression to identify the key contributors to performance degradation. Since DSC values fell between 0 and 1 and that the DSC distributions of the predicted contours for the prostate, rectum and bladder were left skewed, we applied a logit transformation to the DSC values. Subsequently, these logit-transformed DSC values were fitted into a linear model that incorporated the variables mentioned above and their second-order interactions. We preset a significance level at 0.05 to determine the statistical relevance of each factor. All the analyses were performed using SAS 9.4.

## 3. Results

### 3.1. Model performance deterioration

Figures 2(a)–(c) illustrate the temporal trends in contour quality (DSC) for the prostate, rectum, and bladder. We observed distinct performance patterns for each organ. While the quality of prostate and rectum contours declined significantly over time, bladder contours remained relatively stable.




**Figure 2.** Trends in auto-generated contour quality. (a), (b), and (c) present the exponential weighted moving average (EMA) of dice similarity coefficient (EMA DSC) over time post-simulated model deployment for: (a) prostate EMA DSC, (b) rectum EMA DSC, and (c). bladder EMA DSC. EMA DSC for the auto-generated prostate and rectum contours declined, but those for the bladder contours remained stable.

The model's performance trajectory can be divided into three periods:

- Year 2012–14: Following its initial deployment, the model exhibited peak performance for all three organs.
- Year 2015–19: A marked decline in contour quality for the prostate and rectum.
- Year 2020–22: Modest improvements noted in the quality of the prostate and rectum contours; however, not reaching the quality at initial deployment of the model (2012–14).

As shown in table 1, the results from the years 2012–14 in the testing dataset align closely with those of our validation data regarding the quality of auto-generated contours for the prostate and rectum. Between 2012–14 and 2015–19, the median prostate contour DSC reduced from 0.85 to 0.78 ($p < 0.001$), rectum contour DSC reduced from 0.83 to 0.80 ($p < 0.001$). The bladder contour exhibited a slight decrease in median DSC from 0.96 to 0.95 ($p = 0.003$).

Inspection of the EMA DSC curves corroborated these findings. The prostate's EMA DSC exhibited a pronounced decline—0.84–0.74 between 2015 and 2019. Concurrently, the rectum's EMA DSC decreased from 0.82 to 0.76.

### 3.2. Data drift

From 2006 to 2011, the majority of patients underwent conventionally fractionated radiotherapy, with a small subset (17.8%) participating in an SBRT clinical trial. From 2012 to 2014, conventionally fractionated radiotherapy remained prevalent, but a significant transition to SBRT was observed post-2015. SBRT, characterized by its precision facilitated by specialized equipment and refined imaging, predominantly utilizes CT scans with a 2 mm slice thickness, compared to the 3 mm thickness common in conventional treatments. The incorporation of hydrogel spacers and IV contrast became more prevalent in SBRT protocols with hydrogel spacers also being used in some patients undergoing moderately hypofractionated treatments. Institutionally, if feasible, MRI-guided prostate contouring was the preferred method for most patients, especially after 2015. As precision radiotherapy gained prominence, these techniques have become foundational in most radiotherapy planning protocols for prostate cancer patients.

In our clinical practice, the contours for the target and organs at risk (OARs) are delineated on CT images. Over the years, the scanning parameters have remained stable at our institute. Nearly all patients (99.4%) were scanned using a Philips CT simulator, with a tube voltage set at 120 kVp (99.2%) and dynamic tube current modulation for optimal exposure control for prostate cancer patients.

Figure 3 presents a comprehensive overview of the data distribution across different time periods, highlighting significant changes over the years. Notably, CT scan slice thickness distribution shifted from a mix of 1.5 mm ($n = 31$, 19.0%), 2 mm ($n = 32$, 19.6%), and 3 mm ($n = 100$, 61.3%) during 2006–11, to predominantly 2 mm ($n = 236$, 65.6%) and 3 mm ($n = 124$, 34.4%) from 2020–22. Hydrogel spacer and MRI-guided contouring utilizations were absent in the initial period but were adopted by 84.4% ($n = 304$) and 92.5% ($n = 333$) of patients respectively, during 2020–22. IV contrast usage rose from 6.7% ($n = 11$) to 41.9% ($n = 151$). A shift in clinical personnel was observed, with Physician A treating 84.7% ($n = 138$) of patients initially, but in the later period (2020–22), Physicians B ($n = 128$/35.6%), C ($n = 100$/27.8%), and D ($n = 120$/33.3 %), collectively treated 96.7% of cases. All the shifts reflected the evolution in treatment approaches and team composition over the years.

**Table 1.** Results for auto-generated contour quality.

| | | | Hydrogel spacers | | | Slice thickness | | MRI | | IV contrast | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Year | Overall | W/O | Type I | Type II | 3 mm | 2 mm | W/O | W/ | W/O | W |
| Prostate median DSC (Q1, Q3)[a] | Validation[b] | 0.86 (.84, .88) | — | — | — | — | — | — | — | — | — |
| | 2012–14 | 0.85 (.82, .87) | 0.85 (.82, .87) | — | — | 0.85 (.82, .87) | 0.79 (.77, .82) | 0.85 (.82, .87) | 0.84 (.81, .87) | 0.85 (.82, .87) | 0.85 (.77, .87) |
| | 2015–19 | 0.78 (.72, .82) | 0.81 (.77, .85) | 0.76 (.71, .81) | — | 0.81 (.74, .85) | 0.77 (.71, .81) | 0.81 (.76, .85) | 0.77 (.71, .81) | 0.78 (.70, .82) | 0.78 (.75, .82) |
| | 2020–22 | 0.77 (.73, .81) | 0.76 (.72, .82) | 0.75 (.71, .80) | 0.78 (.74, .82) | 0.77 (.71, .82) | 0.77 (.73, .81) | 0.77 (.74, .83) | 0.77 (.73, .81) | 0.76 (.72, .80) | 0.78 (.74, .82) |
| | Test all (2012–22) | 0.78 (.72, .82) | 0.84 (.79, .86) | 0.76 (.71, .81) | 0.78 (.74, .82) | 0.83 (.77, .86) | 0.77 (.72, .81) | 0.84 (.80, .86) | 0.77 (.72, .81) | 0.80 (.73, .85) | 0.78 (.75, .82) |
| Rectum median DSC (Q1, Q3) [a] | Validation[b] | 0.84 (.76, .87) | — | — | — | — | — | — | — | — | — |
| | 2012–14 | 0.83 (.79, .87) | 0.83 (.79, .87) | — | — | 0.83 (.79, .87) | 0.72 (.69, .75) | 0.83 (.78, .87) | 0.83 (.81, .87) | 0.83 (.79, .87) | 0.82 (.79, .84) |
| | 2015–19 | 0.80 (.74, .84) | 0.82 (.78, .85) | 0.79 (.74, .83) | — | 0.81 (.77, .85) | 0.79 (.73, .83) | 0.82 (.77, .85) | 0.79 (.74, .83) | 0.80 (.75, .84) | 0.80 (.74, .84) |
| | 2020–22 | 0.80 (.74, .84) | 0.79 (.73, .84) | 0.80 (.74, .84) | 0.80 (.75, .84) | 0.81 (.76, .85) | 0.79 (.74, .84) | 0.82 (.73,.87) | 0.80 (.74, .84) | 0.79 (.74, .84) | 0.81 (.75, .84) |
| | Test all (2012–22) | 0.80 (.74, .84) | 0.83 (.78, .86) | 0.79 (.74, .83) | 0.80 (.75, .84) | 0.82 (.77, .86) | 0.79 (.74, .83) | 0.83 (.77, .86) | 0.79 (.74, .84) | 0.81 (.75, .85) | 0.81 (.75, .84) |
| Bladder median DSC (Q1, Q3) [a] | Validation[b] | 0.95 (.93, .96) | — | — | — | — | — | — | — | — | — |
| | 2012–14 | 0.96 (.94, .96) | 0.96 (.94, .96) | — | — | 0.96 (.94, .96) | 0.93 (.92, 0.95) | 0.96 (.94, .96) | 0.96 (.95, .97) | 0.96 (.94, .96) | 0.93 (.91, .95) |
| | 2015–19 | 0.95 (.93, .96) | 0.95 (.94, .96) | 0.95 (.93, .96) | — | 0.95 (.93, .96) | 0.95 (.93, .96) | 0.95 (.94, .96) | 0.95 (.93, .96) | 0.95 (.94, .96) | 0.95 (.93, .96) |
| | 2020–22 | 0.95 (.94, .96) | 0.95 (.93, .96) | 0.95 (.94, .96) | 0.95 (.94, .96) | 0.95 (.94, .96) | 0.95 (.94, .96) | 0.94 (.92, .96) | 0.95 (.94, .96) | 0.95 (.94, .96) | 0.95 (.94, .96) |
| | Test all (2012–22) | 0.95 (.94, .96) | 0.95 (.94, .96) | 0.95 (.93, .96) | 0.95 (.94, .96) | 0.95 (.94, .96) | 0.95 (.94, .96) | 0.95 (.94, .96) | 0.95 (.94, .96) | 0.95 (.94, .96) | 0.95 (.93, .96) |

[a] Q1 the first quantile or 25th percentile of the DSC values; Q3 the third quantile or 75th percentile of the DSC values.

[b] We had 33 cases in our validation data set, with specific characteristics: none involved Hydrogel Spacers or MRI-guided contouring; one case involved IV Contrast; slice thicknesses were 3 mm for 18 cases, 2 mm for 10 cases, and 1.5 mm for 5 cases. Due to the small sample sizes in these subgroups, we presented only the overall validation results.

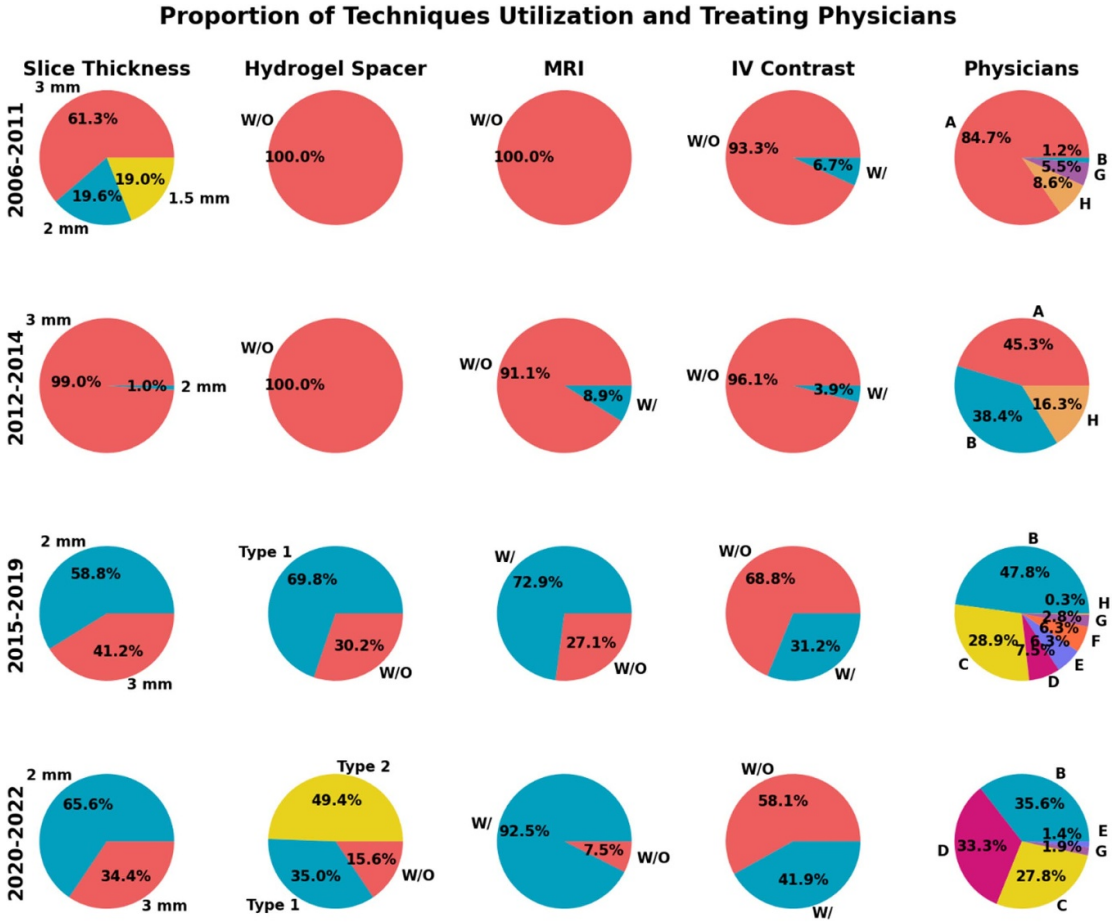



**Figure 3.** Trends in data distribution. Distribution of treatment techniques and physician involvement for the cohorts during year 2006–11, 2012–14, 2015–19, and 2020–22.

### 3.3. Factors contributing to the performance deterioration

*3.3.1. Impact factors to prostate contour performance deterioration*

Figure 4(a) depicts a decline in the model's prostate contour quality, correlating with the increased incorporation of diverse clinical parameters, including various hydrogel spacer types, slice thickness variations, MRI-guided contouring, and IV contrast usage starting from 2015.

Figure 4(b) and table 1 collectively indicate superior prostate contour quality during the 2012–14 when techniques similar to those used in 2006–11. Specifically, absence of spacers yielded the highest overall median DSC of 0.84, surpassing type II (0.78) and type I (0.76) spacer groups. A 3 mm slice thickness group exhibited a median DSC of 0.83, outperforming the 2 mm group (0.77). In the context of MRI-guided contouring, the group without this technique registered a superior median DSC of 0.84, compared to the MRI-guided contouring group, which recorded a median DSC of 0.77. Contrastingly, the performance of the IV contrast groups diverged notably during 2020–22, with the IV contrast group achieving a median DSC of 0.78, surpassing the non-contrast group's 0.76. Additionally, the trend of narrowing differences in subgroup performances suggests potential interaction effects among these variables.

Figure 4(c) delineates physician-stratified EMA DSC curves, revealing consistent declining trends across subgroups. Physician A's contours were notably superior, establishing a contour quality hierarchy across the evaluated periods: during 2012–14: A > B; during 2015–19: B > C and E; during 2020–22: B > E > C > D.

The multiple linear regression analysis revealed significant associations between prostate contour quality and factors including physician contouring styles, hydrogel spacer types, CT scan slice thickness, MRI-guided contouring, and IV contrast use ($p < 0.0001$, $p < 0.0001$, $p = 0.0085$, $p = 0.0012$, $p < 0.0001$, respectively, for the main effects only model). We also identified significant interactions between slice thickness and hydrogel spacer types ($p = 0.0026$), as well as interactions between physicians' styles and MRI-guided contouring ($p = 0.0115$), which impacted prostate contour quality.

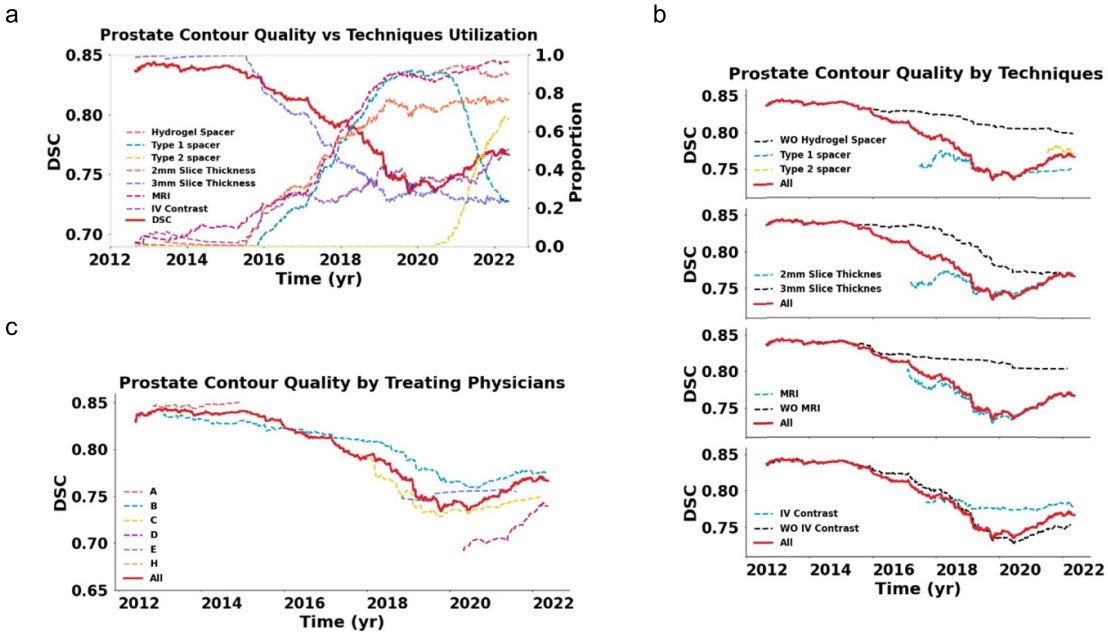



**Figure 4.** Auto-generated prostate contour quality vs. techniques utilization and subgroups. (a) EMA DSC for prostate contour quality and proportion of patients with different techniques utilization versus time after the simulated model deployment. (b) Prostate contour quality for subgroups stratified by techniques used versus time. (c) Prostate contour quality for subgroups stratified by treating physicians versus time. The minimum number of observations for the subgroup EMA DSC curves is set to be 30 in figure (b) due to small sample size of patients treated by some physicians. Physicians with available prostate contours fewer than 30 will not have EMA DSC curves.

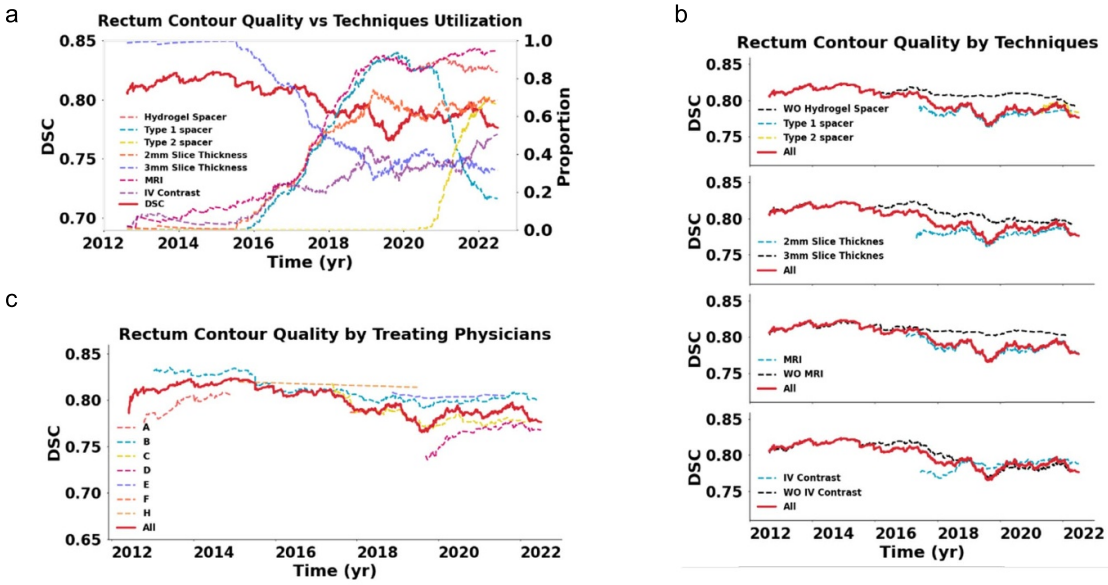



**Figure 5.** Auto-generated rectum contour quality vs. techniques utilization and subgroups. (a) EMA DSC for rectum contour quality and proportion of patients with different techniques utilization versus time after the simulated model deployment. (b) Rectum contour quality for subgroups stratified by techniques used versus time. (c) Rectum contour quality for subgroups stratified by treating physicians versus time. The minimum number of observations for the subgroup EMA DSC curves is set to be 30 in figure (c) due to small sample size of patients treated by some physicians. Physicians with available rectum contours fewer than 30 will not have EMA DSC curves.

### *3.3.2. Impact factors to rectum contour performance deterioration*

Figure 5 illustrates trends for rectum contour quality and its subgroups. The rectum contour quality had similar but less pronounced discrepancies for subgroup EMA DSC curves stratified by hydrogel spacer, MRI-guided contouring, slice thickness or IV contrast compared with those of the prostate. Additionally, the rectum contour quality for subgroups maintaining consistent techniques also decreased, indicating no single variable could fully explain the reduction in contour quality.

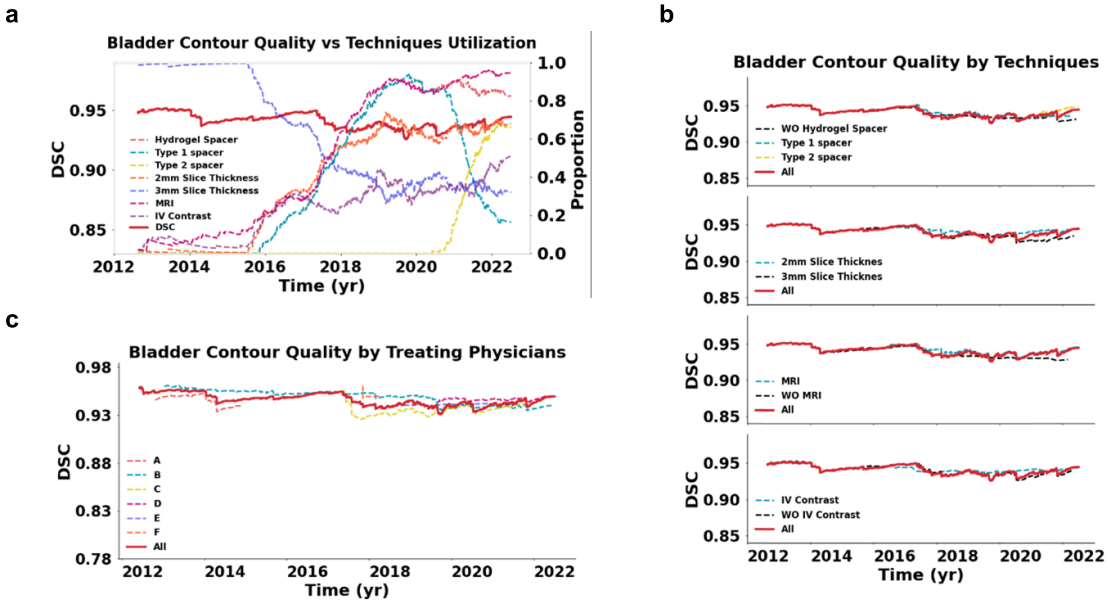



**Figure 6.** Auto-generated bladder contour quality vs. techniques utilization and Subgroups. (a) EMA DSC for bladder contour quality and proportion of patients with different techniques utilization versus time after the simulated model deployment. (b) Bladder contour quality for subgroups stratified by techniques used versus time. (c) Bladder contour quality for subgroups stratified by treating physicians versus time. The minimum number of observations for the subgroup EMA DSC curves is set to be 30 in figure (c) due to small sample size of patients treated by some physicians. Physicians with available bladder contours fewer than 30 will not have EMA DSC curves.

Figure 5(c) suggests the model demonstrated superior accuracy for physician B compared to other physicians for generating rectum contours. The contour quality hierarchy was: during 2012–14: B > H > A; during 2015–19: H > E > B > C and F; during 2020–22: E > B > C > D.

The multiple linear regression analysis revealed that rectum contour quality was significantly influenced by factors including slice thickness, physicians' styles, and the use of different types of hydrogel spacers, with $p$-values of $<0.0001$, $<0.0001$, and 0.0470, respectively.

*3.3.3. Impact factors to bladder contour performance deterioration*

Figure 6 underscores the bladder contour quality's stability, demonstrating its resilience to changes in clinical techniques and personnel. The EMA DSC curves of different subgroups are nearly indistinguishable, each maintaining a median DSC of 0.95, indicating consistent performance. However, as shown in table 1, a deviation is noted between 2012 and 2014 where the IV contrast group exhibited a lower median DSC (0.93) compared to the non-contrast group (0.96). Meanwhile, the model showcased consistent efficacy across all physicians (figure 6(c)) with minimal variations.

The outcomes of the multiple linear regression analysis indicated the bladder contour quality was primarily influenced by the IV contrast, with a $p$-value of 0.055.

### 3.4. Model updates

We found the utilization of the hydrogel spacers with different types, different slice thicknesses, MRI-guided contouring, IV contrast and different physicians' contouring styles are contributors to the DL-based auto-segmentation model's performance deterioration.

To counteract this deterioration, we instigated model updates, incorporating enriched data sets to enhance performance. Specific updates were triggered upon the identification of discernible performance declines and conducted every two years.

- Update Model 1 was trained on data spanning 2006–16 and tested on data from 2017–22.
- Update Model 2 utilized training data from 2006–18 and tested on 2019–22 data.
- Update Model 3 was trained using data from 2006–20 and tested on 2021–22 data.

Figures 7(a)–(c) delineate the EMA DSC curves for prostate, rectum, and bladder contours, respectively, illustrating performance enhancements post-update. Each model update manifested in improved contour predictions, although we noticed decline in rectum contour quality in Update Model 2 relative to Model 1 (figure 7(b)) in 2019, which was attributed to disparate initial values influencing the EMA curves. Post-2019,

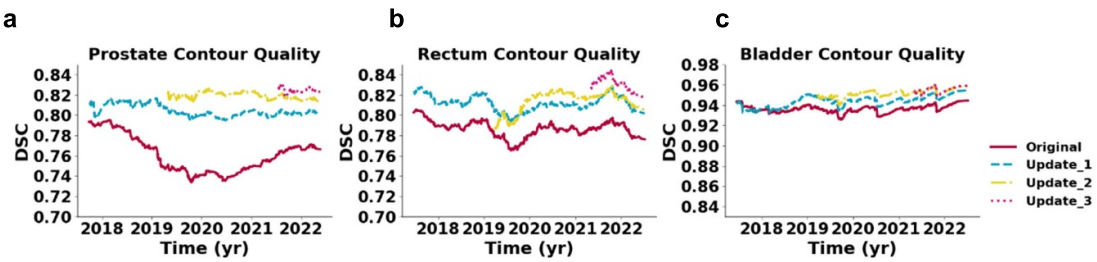



**Figure 7.** Quality of the model generated contour (EMA DSC) for different models. (a) prostate EMA DSC, (b) rectum EMA DSC, and (b) bladder EMA DSC. The first and second updated model by adding shifted data, the permance for predicting prostate, rectum and bladder contours improved. The third update also improved the model permance for predicting prostate and rectum contour quality.

the EMA DSC curve for Update Model 1 benefited from the weighted averaging effect of preceding higher values, which was not extended to Update Model 2.

## 4. Discussion and conclusions

In this simulated yet practical study evaluating the performance of a DL model deployed in the clinic, we showed that performance decreased over time. We then investigated and identified the potential factors contributing to this deterioration, explicitly evaluating the introduction of new technologies and procedures like hydrogel spacer, CT slice thickness, MRI-guided contouring and the presence of contrast agent in the bladder. Additionally, we considered the changing clinical personnel. To the best of our knowledge, no other studies in the DL auto-contouring medical domain have undertaken a similar exploration into identifying and analyzing factors contributing to the performance deterioration of a deployed DL model over time.

Many DL models published in the literature demonstrate the clinical utility on their specific training and testing datasets [29–34]. However, many of these studies do not evaluate the models' performance years later. Given the evolving of medical practices, shifts in treatment paradigms are inevitable. Such changes can pose challenges for DL models that are not periodically updated for clinical application. Our research revealed our model, trained on data from 2006 to 2011, maintained clinically acceptable performance until the end of 2014. After this period, its ability to contour the prostate and rectum declined significantly. This decline is concerning, especially considering the critical role these structures play in prostate cancer radiotherapy. Radiotherapy targets the prostate while trying to spare adjacent structures, like the rectum and bladder, from receiving toxic radiation doses [35]. An inappropriately contoured target can lead to poor treatment coverage [36], which increases the patient's risk for recurrence or increases the patient's risk for toxicity if the contoured target inadvertently includes normal tissues, such as the rectum.

It is critical to understand why models might fail in the future. Thus we explored several potential factors that could have impacted our model's performance: the evolution of new treatments and changing personnel. Prostate cancer treatment has evolved over the last several decades. Initially, patients were treated with conventionally fractionated radiotherapy, which included daily radiation treatment, five days a week for more than seven weeks [37]. Over more recent years, patients are more likely to be treated with hypofractionated approaches, which may take only two to four weeks [38]. However, early studies for these hypofractionated approaches did not arise until 2007 [39], with their mainstream clinical adoption occurring much later.

Given this shift, we investigated potential influence of SBRT on our model. Specifically, we assessed the impact of various factors commonly associated with SBRT treatment, such as CT slice thickness, hydrogel spacer usage, MRI-guided contouring, and the use of contrast. Our findings revealed that subgroups utilizing hydrogel spacers (type I and type II), MRI-guided contouring, or 2 mm thick CT slices generally exhibited lower contour quality compared to their counterparts. Specifically, the model's predictions for prostate and rectum contours were inferior in these subgroups compared to those without hydrogel spacers, without MRI-guided contouring, or with 3 mm thick CT slices. Interestingly, the presence of contrast in the bladder enhanced the model's accuracy in predicting prostate contours. However, it is crucial to note that no single technique was solely responsible for the model's performance decline. This is clearly illustrated by the marked decreases observed in the 'without certain technique group' EMA DSC curves (e.g. WO Hydrogel Spacer in figure 4(b)), underscoring the multifactorial nature of the performance deterioration.

We quickly observed that the model's ability to contour the rectum and the prostate declined after the type I hydrogel spacer's introduction into the clinic. However, the prostate contour quality improved significantly after type II hydrogel spacer's use in the clinic. In contrast, bladder contour quality remained

relatively stable, primarily influenced by the use of IV contrast. Clinically, a hydrogel creates a gap between the prostate and the rectum, reducing rectal radiation exposure and associated side effects [40]. We believe this introduced a data distribution shift by systematically altering traditional anatomy, given that the original model was trained before the FDA's approval of the hydrogel spacer in 2015 [41].

The type I hydrogel spacer, while indistinguishable on CT scans, was visible on MRI images. Consequently, contours for SBRT cases, which predominantly used hydrogel spacers, were delineated with MRI guidance, adhering to our institutional protocol. In contrast, the brighter type II hydrogel spacer was evident on CT scans due to its higher Hounsfield Unit. The pre-trained DL model likely interpreted the 'brighter' type II hydrogel spacer as non-prostate tissue, leading to a noticeable gap between the auto-generated prostate and rectum contours for patients with type II spacer. We speculate that the effect of contrast present within the bladder mirrors that of the type II hydrogel spacer, aiding the model in distinguishing non-prostate tissues. This could explain the enhanced prostate contour quality in patients with contrast present in the bladder. A similar rationale applies to the prediction of rectum contours in the presence of hydrogel spacers and bladder contrast, though the impact of the latter was not statistically significant.

We also found MRI-guided contouring and physicians' contouring styles correlated with the model performance in predicting the prostate and rectum contours. Notably, an interaction effect between physicians' contouring styles and MRI-guided contouring was observed for the prostate contour quality. This suggests that physicians' contours vary depending on whether they utilize MRI imaging or not.

The exact reasoning is unknown, but we postulate that physicians are trained to prioritize target coverage and often contour the entire target based on their expertise and interpretation. While there are established guidelines for contouring the prostate and rectum, the delineation can be ambiguous in certain imaging modalities. MRI-guided contouring, in contrast to CT imaging alone, offers enhanced accuracy and potentially more consistent target and OARs contours [42]. This precision can lead to variations in contouring, especially when combined with individual physician preferences.

Furthermore, it is plausible that the introduction of MRI-guided contouring has influenced physicians to adapt or modify their contouring techniques, given the superior soft tissue contrast and detail provided by MRI. This adaptation might result in contours that deviate from those based solely on CT, leading to discrepancies when compared with DL model predictions trained on CT-based contours.

Additionally, while not widespread, there might be instances where physicians defer OARs contouring to residents or other staff, making only minor adjustments themselves. Such delegation can introduce variability, especially if the training or experience level of the delegates differ from the physicians.

The utilization of a 2 mm thick slice in CT scans, favored for its precision and detailed representation, is predominantly used for SBRT procedures. Notably, the proportion of CT scans employing this 2 mm thickness escalated from 19.6% (training/validation) to 50.8% (testing). However, intriguingly, our observations indicated that the contour quality for both the prostate and rectum was inferior with the 2 mm thick slice compared to the 3 mm thick slice. This counterintuitive finding suggests that while finer slices capture more anatomical details, they might introduce complexities or nuances that challenge the current DL model's ability to generate accurate contours which was trained predominantly on contours based on 3 mm thick CT images.

In this auto-segmentation task, our data consist of input CT images (explanatory variable $X$) and the delineated contours (response variable $Y$). Variations in the joint distribution $P(X,Y)$ indicate a drift in data distribution, which influence model performance. The factors contributing to this study's variability can be divided into changes within the CT images ($P(X)$) and variations in the contours ($P(Y)$).

Changes in $P(X)$ include factors such as variations in CT imaging slice thickness and the use of IV contrast, both of which significantly alter the imaging characteristics. Additionally, the use of hydrogel spacers represents another change $P(X)$ by introducing physical discrepancies in the CT images, compared to scans without hydrogel spacers.

On the other hand, variations in $P(Y)$ encompass individual physician contouring styles and MRI-guided contouring techniques. The latter, which utilizes MRI scans as a reference, directly affects the precision and quality of the delineated contours.

Understanding and addressing these factors are essential for mitigating the impact of data distribution drift on the efficacy of auto-segmentation models in medical imaging. Other possible explanations for the observed decline in the model's performance need further exploration, especially if the model is to be integrated into clinical practice. This observation underscores the critical need for continuous and systematic monitoring of deployed models. We advocate for the establishment of guidelines that methodically oversee these models' performance, coupled with strategies to investigate contributing variables. Leveraging out-of-distribution quantification metrics, such as uncertainty estimation techniques, might offer a case-specific approach to identify testing instances that diverge from the training dataset [43, 44].

## 5. Conclusion

In this novel study, we clearly demonstrated that a model's performance can deteriorate significantly over time. While we could not draw definitive causal conclusions from this retrospective analysis, we pinpointed potential variables that led to shifts in data distribution and subsequent performance decline. Based on these insights, we updated the model biennially using post-deployment data, witnessing consistent performance improvements after each iteration. These findings are not just limited to the specific model and clinical task but are broadly applicable to other models employed in medicine. Thus, it is crucial to establish guidelines for regular monitoring and refinement of these models, ensuring their continued efficacy and contribution to patient care across various medical domains.

## Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

## Acknowledgments

We would like to thank Ms. Sepeadeh Radpour for editing the manuscript.

## Funding

This study is supported by NIH Grants R01CA237269, R01CA254377, and R01CA258987.

## Contributors

Biling Wang and Michael Dohopolski provided data resources, curated data, searched the literatures, wrote and revised the manuscript. Biling Wang, Michael Dohopolski and Steve Jiang had full access to the data and verified the data. Biling Wang trained the deep learning model and analyzed the data. Ti Bai constructed the deep learning framework and guided the model training. Junjie Wu downloaded all the DICOM data and helped to analyze the DICOM data. Dan Nguyen guided the model training and the analysis. Mu-Han Lin and Michael Dohopolski guided the data curation and contributed essential clinical insights. Raquibul Hannan, Neil Desai, Aurelie Garant, Daniel Yang and Robert Timmerman provided data annotations and clinical expertise. Xinlei Wang guided the overall research study, supervised the statistical evaluation, and contributed editorial assistance to the manuscript. Steve Jiang was pivotal in conceptualizing and directing the research, providing overall guidance on the direction and goals of the project, ensuring the integrity and quality of the research, and contributing editorial assistance to the manuscript. All co-authors have reviewed and consented to the published version of the manuscript.

## ORCID iDs

Biling Wang https://orcid.org/0000-0002-6894-8645
Ti Bai https://orcid.org/0000-0002-6697-7434
Dan Nguyen https://orcid.org/0000-0002-9590-0655
Xinlei Wang https://orcid.org/0000-0002-8561-6511
Steve B Jiang https://orcid.org/0000-0002-3083-6752

## References

[1] Fosso Wamba S and Queiroz M M 2021 Responsible artificial intelligence as a secret ingredient for digital health: bibliometric analysis, insights, and research directions *Inf. Syst. Front.* **25** 1–16
[2] Rajpurkar P, Chen E, Banerjee O and Topol E J 2022 AI in health and medicine *Nat. Med.* **28** 31–38
[3] Group D-AS 2021 DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence *Nat. Med.* **27** 186–7
[4] Kundu S 2021 AI in medicine must be explainable *Nat. Med.* **27** 1328
[5] Wang F, Casalino L P and Khullar D 2019 Deep learning in medicine-promise, progress, and challenges *JAMA Intern. Med.* **179** 293–4
[6] Jasper van Hoek H A, Leichtle A, Härmä K, Hilt D, von Tengg-kobligk H, Heverhagen J and Poellinger A 2019 A survey on the future of radiology among radiologists, medical students and surgeons: students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over *Eur. J. Radiol.* **121** 108742
[7] Davenport T K R 2019 The potential for artificial intelligence in healthcare *Future Healthcare J.* **6** 94–98

[8] Young A S 2022 AI in healthcare startups and special challenges *Intell. Med.* **6** 100050
[9] Agency I A E 2013 *Roles and Responsibilities, and Education and Training Requirements for Clinically Qualified Medical Physicists: Internat* (Atomic Energy Agency)
[10] Kawaguchi K, Kaelbling L P and Bengio Y 2017 Generalization in deep learning (arXiv:171005468)
[11] Futoma J, Simons M, Panch T, Doshi-Velez F and Celi L A 2020 The myth of generalisability in clinical research and machine learning in health care *Lancet Digit. Health* **2** e489–e92
[12] Finlayson S G, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, Kohane I S and Saria S 2021 The clinician and dataset shift in artificial intelligence *New Engl. J. Med.* **385** 283
[13] Wong A *et al* 2021 External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients *JAMA Intern. Med.* **181** 1065–70
[14] Van Calster B, Wynants L, Timmerman D, Steyerberg E W and Collins G S 2019 Predictive analytics in health care: how can we know it works? *J. Am. Med. Inform. Assoc.* **26** 1651–4
[15] Wang X, Liang G, Zhang Y, Blanton H, Bessinger Z and Jacobs N 2020 Inconsistent performance of deep learning models on mammogram classification *J. Am. Coll. Radiol.* **17** 796–803
[16] Liang X, Nguyen D and Jiang S B 2020 Generalizability issues with deep learning models in medicine and their potential solutions: illustrated with cone-beam computed tomography (CBCT) to computed tomography (CT) image conversion *Mach. Learn. Sci. Technol.* **2** 015007
[17] Chen C *et al* 2020 Improving the generalizability of convolutional neural network-based segmentation on CMR images *Front. Cardiovasc. Med.* **7** 105
[18] Jiang Y F P, Yak S, Roy D M, Mobahi H, Dziugaite G K, Bengio S, Gunasekar S, Guyon I and Neyshabur B 2020 Neurips 2020 competition: predicting generalization in deep learning (NeurIPS(Neural Information Processing Systems)) (arXiv:2012.07976)
[19] Nguyen D *et al* 2021 Deep learning–based COVID-19 pneumonia classification using chest CT images: model generalizability *Front. Artif. Intell.* **4** 87
[20] Subbaswamy A and Saria S 2019 From development to deployment: dataset shift, causality, and shift-stable models in health AI *Biostatistics* **21** 345–52
[21] Kashyap M, Seneviratne M, Banda J M, Falconer T, Ryu B, Yoo S, Hripcsak G and Shah N H 2020 Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network *J. Am. Med. Inform. Assoc.* **27** 877–83
[22] Davis S E, Lasko T A, Chen G, Siew E D and Matheny M E 2017 Calibration drift in regression and machine learning models for acute kidney injury *J. Am. Med. Inform. Assoc.* **24** 1052–61
[23] Davis S E, Lasko T A, Chen G and Matheny M E Calibration drift among regression and machine learning models for hospital mortality.pdf
[24] Challen R, Denny J, Pitt M, Gompels L, Edwards T and Tsaneva-Atanasova K 2019 Artificial intelligence, bias and clinical safety *BMJ Qual. Saf.* **28** 231–7
[25] Nestor B (ed) 2019 Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks *Machine Learning for Healthcare Conf.* (PMLR)
[26] Lee C S and Lee A Y 2020 Clinical applications of continual learning machine learning *Lancet Digit. Health* **2** e279–e81
[27] Feng J, Phillips R V, Malenica I, Bishara A, Hubbard A E, Celi L A and Pirracchio R 2022 Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare *npj Digit. Med.* **5** 1–9
[28] Ronneberger O, Fischer P and Brox T U-net: convolutional networks for biomedical image segmentation 2015 *Int. Conf. on Medical Image Computing and Computer-assisted Intervention p* 234: 41
[29] Gulshan V *et al* 2016 Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs *JAMA* **316** 2402–10
[30] Esteva A, Kuprel B, Novoa R A, Ko J, Swetter S M, Blau H M and Thrun S 2017 Dermatologist-level classification of skin cancer with deep neural networks *Nature* **542** 115–8
[31] Kermany D S *et al* 2018 Identifying medical diagnoses and treatable diseases by image-based deep learning *Cell* **172** 1122–31. e9
[32] Rajpurkar P *et al* 2020 CheXpedition: investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting (arXiv:200211379)
[33] Bai T *et al* 2021 A proof-of-concept study of artificial intelligence assisted contour revision (arXiv:210713465)
[34] Liang X *et al* 2022 Segmentation by test-time optimization (TTO) for CBCT-based adaptive radiation therapy (arXiv:220203978)
[35] Delobel J B *et al* 2017 Nomogram to predict rectal toxicity following prostate cancer radiotherapy *PLoS One* **12** e0179845
[36] Liu H, Amaloo C, Sintay B and Wiant D 2021 Dosimetric effects due to inter-observer variability of organ contouring when utilizing a knowledge-based planning system for prostate cancer *Int. J. Med. Phys. Clin. Eng. Radiat. Oncol.* **10** 47–58
[37] Zelefsky M J, Chan H, Hunt M, Yamada Y, Shippy A M and Amols H 2006 Long-term outcome of high dose intensity modulated radiation therapy for patients with clinically localized prostate cancer *J. Urol.* **176** 1415–9
[38] Morgan S C *et al* 2018 Hypofractionated radiation therapy for localized prostate cancer: executive summary of an ASTRO, ASCO, and AUA evidence-based guideline *Pract. Radiat. Oncol.* **8** 354–60
[39] Madsen B L, Hsi R A, Pham H T, Fowler J F, Esagui L and Corman J 2007 Stereotactic hypofractionated accurate radiotherapy of the prostate (SHARP), 33.5 Gy in five fractions for localized disease: first clinical trial results *Int. J. Radiat. Oncol. Biol. Phys.* **67** 1099–105
[40] Shore N 2018 Spacers and prostate radiation therapy: what urologists should know *Everyday Urol.—Oncol. Insights* **3** (available at: www.urotoday.com/journal/everyday-urology-oncology-insights/articles/108597-spacers-and-prostate-radiation-therapy-what-urologists-should-know.html)
[41] FDA 2015 *Device Classification under Section 513(f)(2)(De Novo) DEN140030* ed (FDA)
[42] McLaughlin P W, Evans C, Feng M and Narayana V 2010 Radiographic and anatomic basis for prostate contouring errors and methods to improve prostate contouring accuracy *Int. J. Radiat. Oncol. Biol. Phys.* **76** 369–78
[43] Dohopolski M, Chen L, Sher D and Wang J 2020 Predicting lymph node metastasis in patients with oropharyngeal cancer by using a convolutional neural network with associated epistemic and aleatoric uncertainty *Phys. Med. Biol.* **65** 225002
[44] Gal Y and Ghahramani Z (eds) 2016 Dropout as a bayesian approximation: representing model uncertainty in deep learning *Int. Conf. on Machine Learning* (PMLR)