## Appendix 1

### *Imaging data acquisition and preprocessing*

All magnetic resonance imaging (MRI) data from local institutions were acquired using 1.5-T or 3.0-T MRI scanners (Fujian Provincial Cancer Hospital: MAGNETOM Verio, MAGNETOM Skyra, or Trio Tim [Siemens Healthineers, Erlangen Germany]; Discovery MR 750 [GE HealthCare, Chicago, IL, USA]; or Ingenia [Philips Healthcare, Amsterdam, the Netherlands]). The pelvic MR imaging protocol included the following sequences: (I) contrast-enhanced T1-weighted imaging (ceT1WI), (II) T2-weighted imaging (T2WI), and (III) diffusion-weighted imaging (DWI). The T1c sequence was acquired immediately after intravenous administration of a gadolinium-based contrast agent at a dose of 0.1 mmol/kg. All imaging data were collected before the initiation of surgical treatment. The quality of the imaging data was assessed to ensure the absence of patient motion and artifacts. The parameters for the ceT1W were as follows: repetition time (TR), 220–1,900 ms; echo time (TE), 2.3–29 ms; section thickness, 2.0–5.0 mm; interslice spacing, 1.5–2.0 mm; number of excitations (NEX), 1; flip angle (FA), 50–111°; field of view (FOV), 220×192–240×240 mm$^2$; and matrix, 256×162–320×256 mm$^2$. The parameters for the T2W sequence were as follows: TR, 1255–6690 ms; TE, 70–122 ms; section thickness, 2.0–5.0 mm; interslice spacing, 1.5–2.0 mm; NEX, 1; FA, 90–142°; FOV, 220×192–240×240 mm$^2$; and matrix, 320×238–512×512 mm$^2$. The parameters for the fluid-attenuated inversion recovery (FLAIR) sequence were as follows: TR, 3,500–12,000 ms; section thickness, 2.0–5.0 mm; interslice spacing, 1.5–2.0 mm; NEX, 1; FA, 90–150°; FOV, 220×192–240×240 mm$^2$; and matrix, 256×179–256×256 mm$^2$.

Preprocessing for image standardization included the entire gross tumor volume cropping, the cropping of all images to a size of 32×128×128 based on the location of the tumor region, and the resampling of anisotropic voxels into 1×1×1 mm$^3$ via linear interpolation.

## Appendix 2

### *Detailed architecture of the segmentation model*

*Figure S1* shows the architecture of the convolution neural network (CNN) encoder, decoder, and feed-forward network in the Transformer. It consists of a convolutional batch normalization leaky rectified linear unit (ReLU), and three stages of 3D residual blocks. The decoder contains four upsampling modules. Each of the first three modules has a "ConvTranspose3d" layer followed by a residual block and a pixel-wise summation with the corresponding feature maps from the encoder and the "ConvTranspose3d" layer. The last module comprises an upsampling layer followed by a 3D convolutional layer that maps the 64-channel feature maps to the desired number of classes. The feed-forward network in the Transformer has two linear projection layers: a Gaussian Error Linear Unit (GELU) activation layer and a dropout layer, with a normalization layer following the first layer and a dropout layer following the second layer.

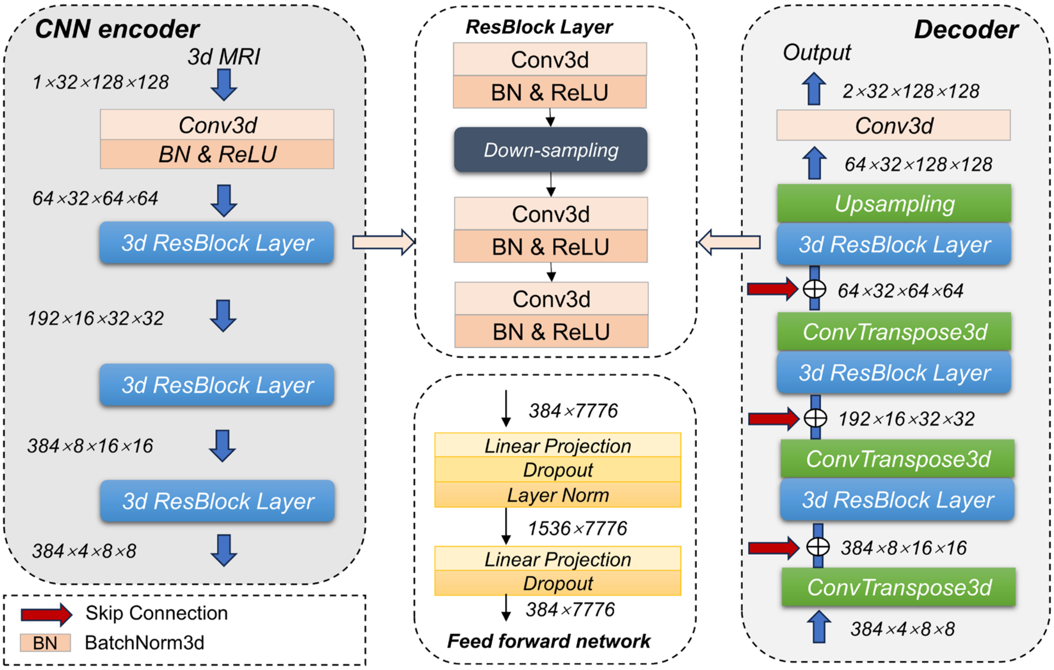


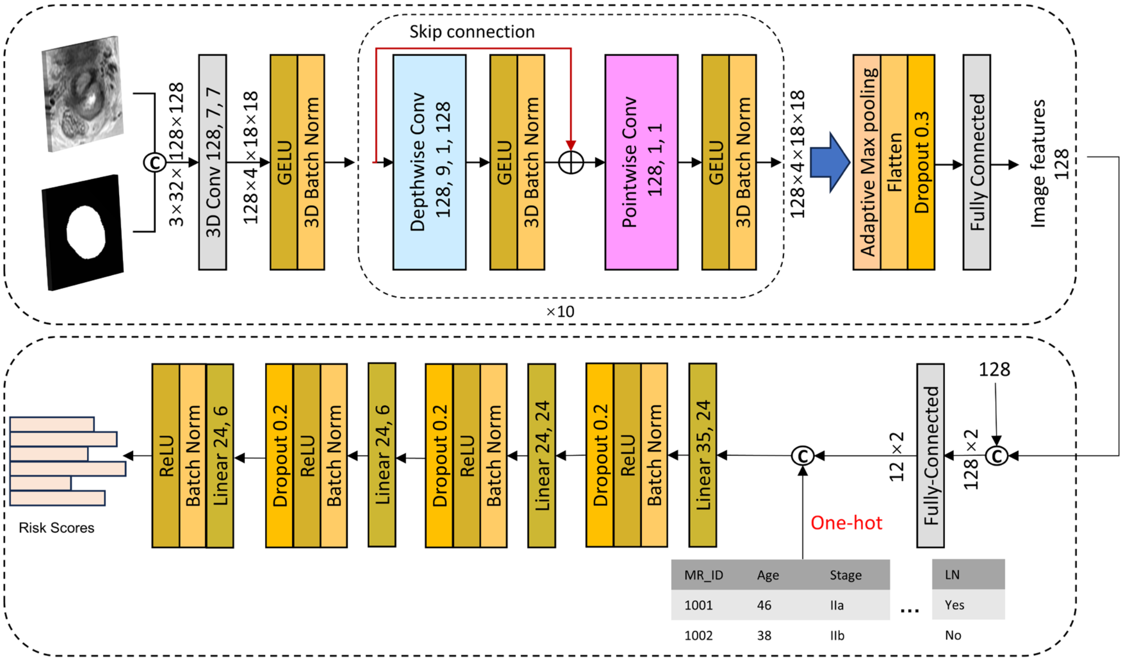
**Figure S1** Network architecture of the CoTr model. CNN, convolution neural network; Conv3d, 3d convolution layer; MRI, magnetic resonance imaging;

## Appendix 3

### *Detailed architecture of the prognostic model*

*Figure S2* shows a detailed schematic of the prognostic architecture. The upper part of the figure is mainly used for extracting the multiscale features of the image. The features are extracted by feeding the MRI scan along with the corresponding mask into the network and finally through a lightweight CNN network. The CNN network is basically a ConvMixer architecture, which uses a mixture of separation space and channel dimensions. It starts with 3D convolution, followed by a ConvMixer structure consisting of a depthwise convolution combined with residual concatenation and a pointwise convolution, which has 10 layers. Each convolution is followed by an activation function GELU and a batch normalization layer. Finally, the features are output after a fully connected layer.

Survival prediction is shown in the lower half of the figure. By combining multiscale image features and clinical information, the model goes through four fully-connected layers, each of which is followed by a batch normalization layer, an activation layer ReLU, and a dropout layer.



| MR_ID | Age | Stage | | LN |
|---|---|---|---|---|
| 1001 | 46 | IIa | ... | Yes |
| 1002 | 38 | IIb | | No |

**Figure S2** Network architecture of the survival prediction model.

## Appendix 4

### *Evaluation metrics*

(I) Dice similarity coefficient (DSC)

$$DSC = \frac{2\left|A \cap B\right|}{\left|A\right| + \left|B\right|} = \frac{2TP}{2TP + FP + FN} \qquad [1]$$

A: Ground truth; B: Segmentation

(II) 95% Hausdorff distance (HD95)

$$HD\left(A, B\right) = \max(h\left(A, B\right), h(B, A)) \qquad [2]$$

In the actual calculation, instead of selecting distances that are not the maximum distance, we take distances ranked at 5% after ranking them from largest to smallest. The purpose of doing this is to exclude some unreasonable distances caused by outliers and to maintain the stability of the overall value.

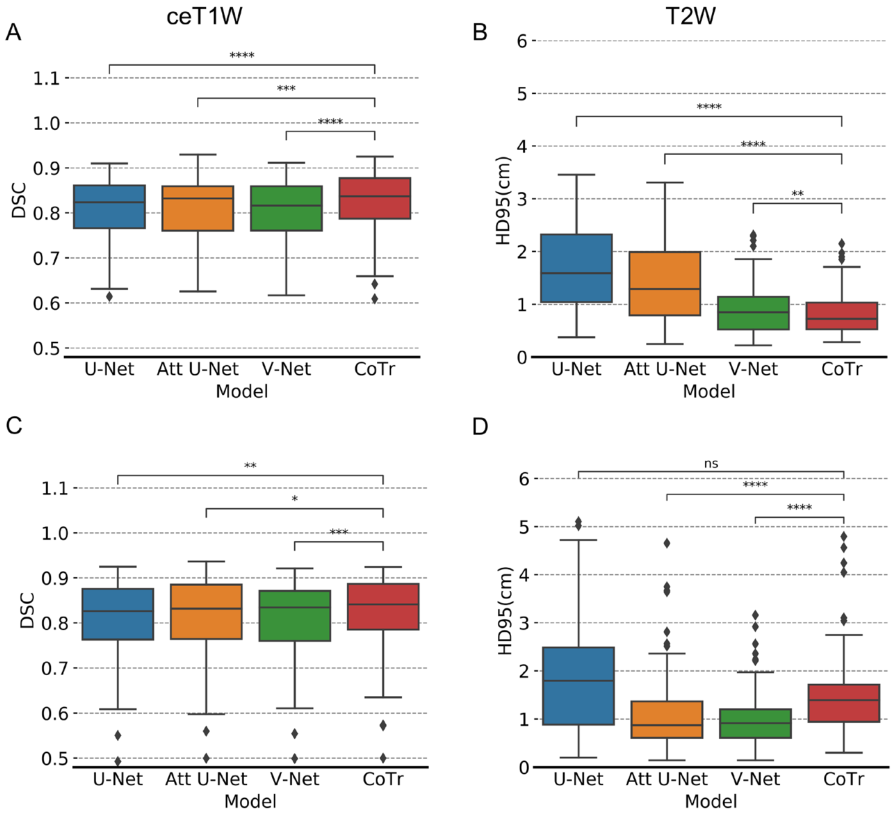


**Figure S3** Boxplots illustrating the DSC and HD95 of each model on the test cohort. (A,B) Comparison of the DSC and HD95 on ceT1WI. (C,D) Comparison of the DSC and HD95 on T2WI. Wilcoxon rank test: *, P<0.05; **, P<0.01; ***, P<0.001; ****, P<0.0001. ceT1W, contrast-enhanced T1-weighted; T2W, T2-weighted; DSC, Dice similarity coefficient; HD95, 95% Hausdorff distance; ns, not significant; Att U-Net, attention U-Net.