Machine learning for prognostic modelling in head and neck cancer using multimodal data

Lessons from the RADCURE Prognostic Modelling Challenge

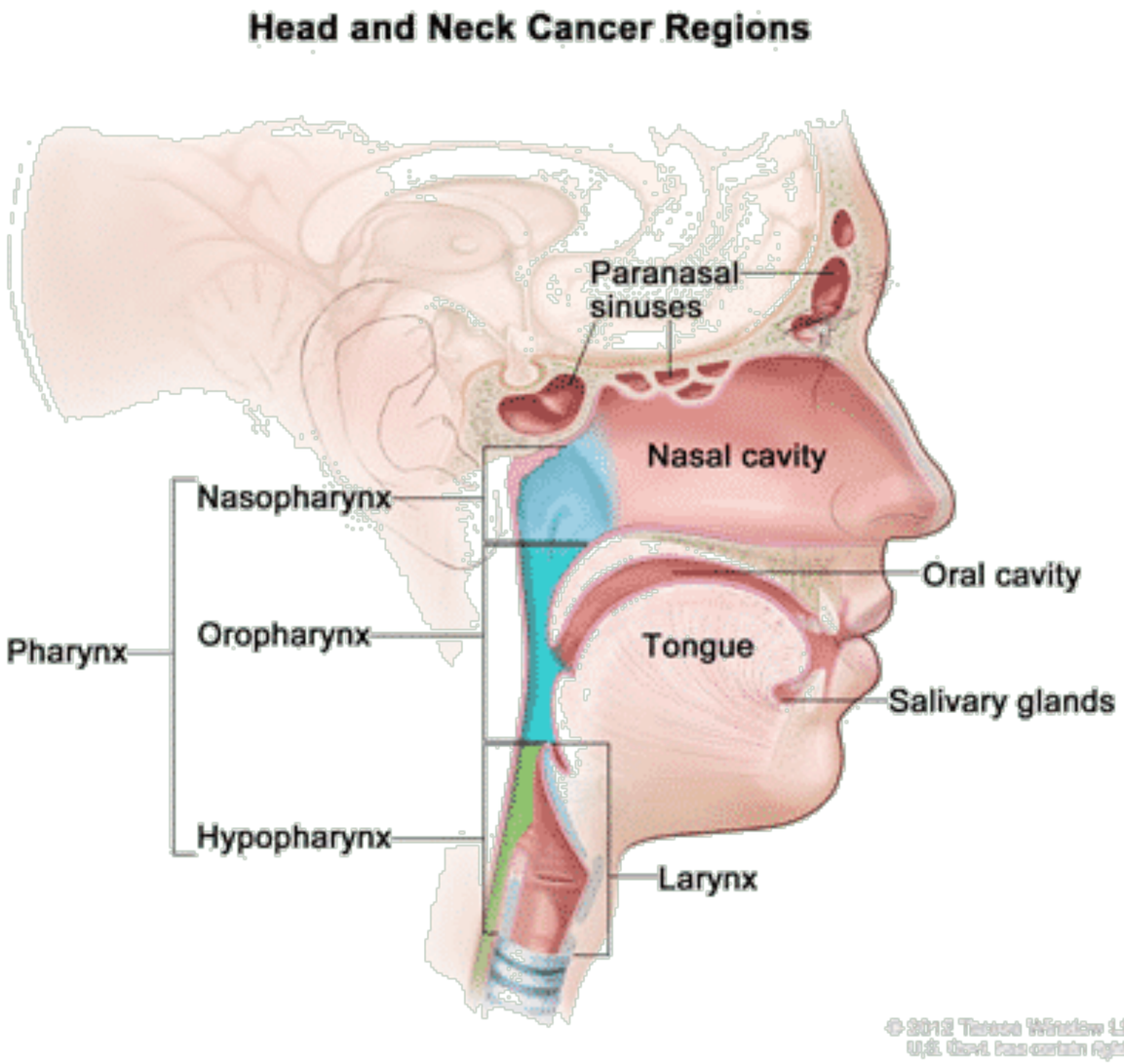
Michal Kazmierski

michal.kazmierski@uhnresearch.ca

Haibe-Kains Lab
Department of Medical Biophysics, University of Toronto
Princess Margaret Cancer Centre

Head and neck cancer

- **Head & neck cancer (HNC)** affects **5.5 million** people per year worldwide, **50% 5 year survival rates**



https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet (accessed 29/03/20)

Head and neck cancer

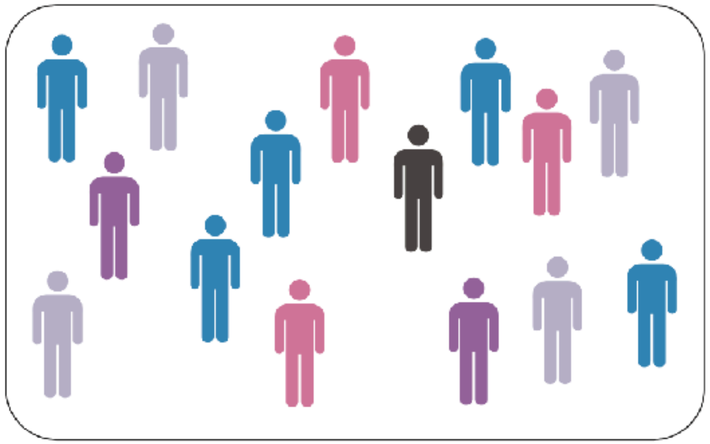
- **Head & neck cancer (HNC)** affects **5.5 million** people per year worldwide, **50% 5 year survival rates**

- **High heterogeneity** in tumour characteristics and prognosis

➡ difficult to individualize treatment

➡ suboptimal treatment outcomes

➡ **need for better prognostic tools to enable personalized treatment and guide clinical decision making**



- Patients with the same tumour disease and stage have typically received similar treatments

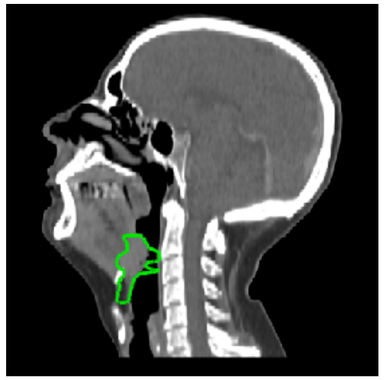
- Biomarkers allow stratification into small subgroups
- Trials for treatment individualization

Baumann et al. Nat Rev Cancer (2016)

Radiomics: emerging prognostic tool for HNC

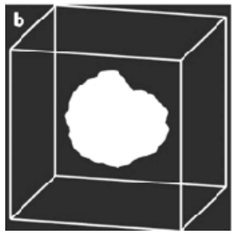
- **Radiomics** uses computational methods to analyze radiological images of tumours, e.g. computed tomography (CT)

➡ non-invasive way to study the tumour *in vivo*

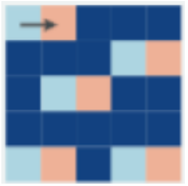


Imaging

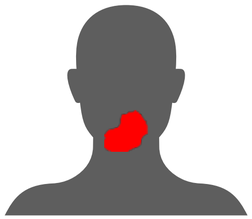
Intensity distribution



Morphology

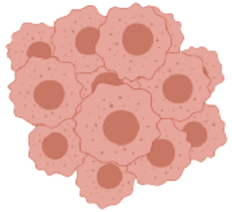


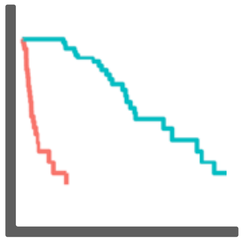
Texture



Anatomy and environment

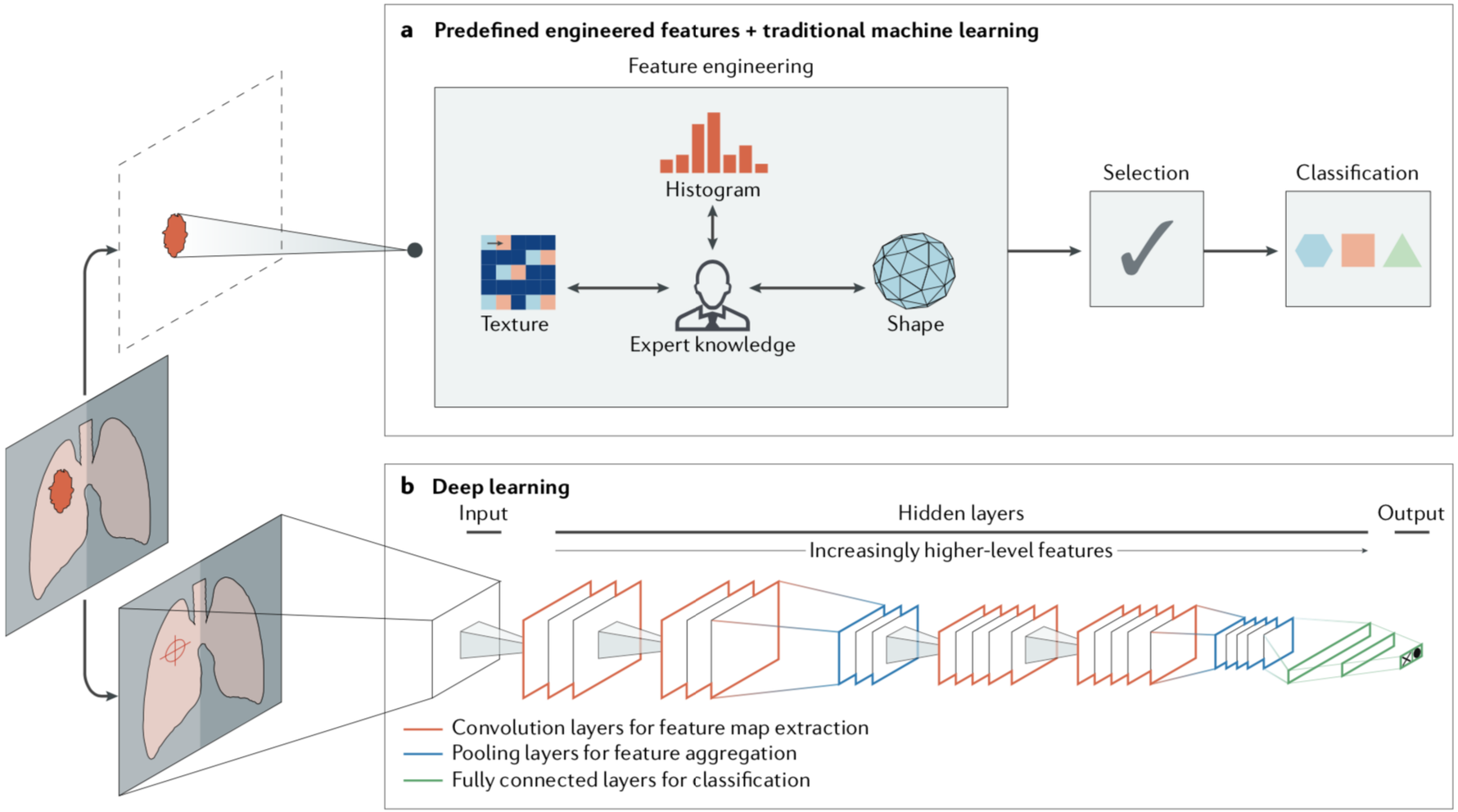
Quantify image characteristics





Link to tumour biology, predict patient outcomes

How to extract information from images?



Engineered — try to manually design image features

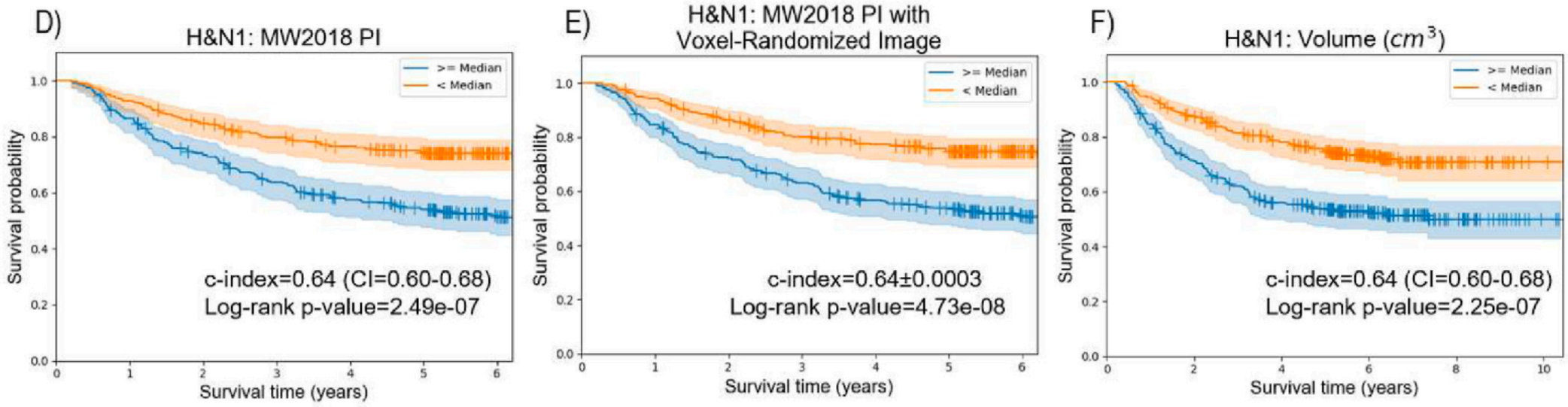
Deep learning — use a ML algorithm to learn the most representative features given labelled training data

Convolutional neural network (convnet): a machine learning (ML) algorithm used for image processing tasks

Hosny et al. Nat Rev Cancer (2018)

CT radiomics for prognosis: the unfulfilled promise?

- Many retrospective studies but limited clinical adoption:
  - Variable **study design, reporting and reproducibility**
  - **Small sample sizes**
  - Lack of **standardized benchmarks:** difficult to compare different approaches
  - Confounding with tumour volume (engineered features)



Welch et al. Radiother Oncol (2019)

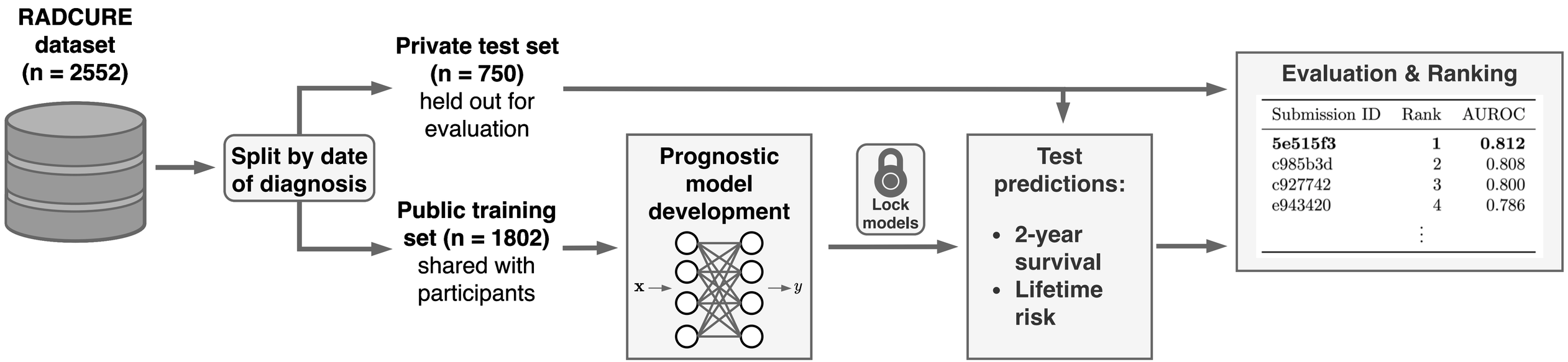
Aims

1. Develop an accurate prognostic machine learning model for HNC survival using routinely collected data from electronic medical records (EMR) and pretreatment CT images

2. Evaluate the true added value of CT radiomics (both engineered and deep learning) compared to other prognostic factors in a reproducible and rigorous way

RADCURE Prognostic Modelling Challenge

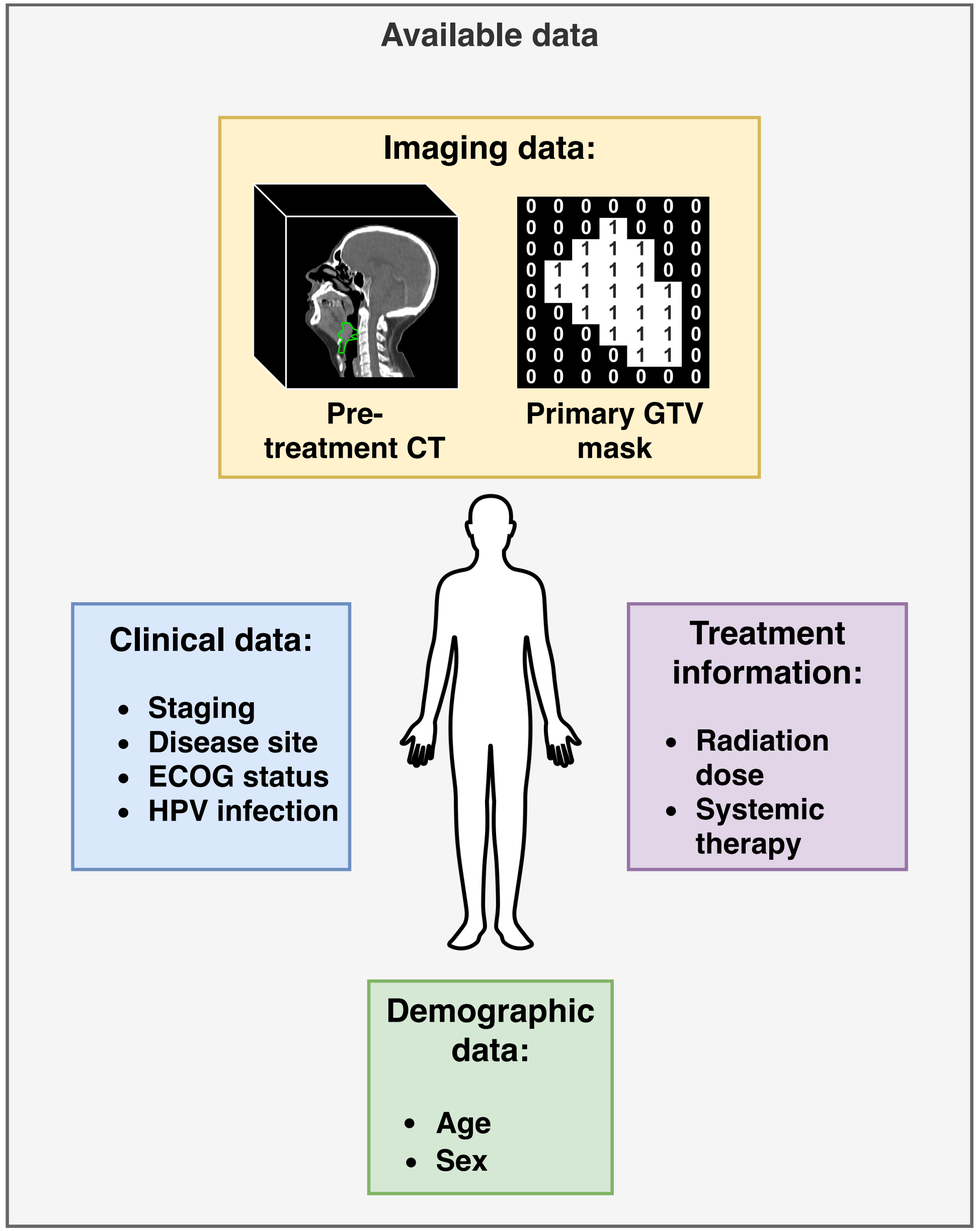
- Conducted by the Princess Margaret Radiomics group between May–September 2020
- Open to teams within the University Health Network system



Dataset

- 2552 HNC patients treated with radio(chemo)therapy at Princess Margaret Cancer Centre between 2005–2018, 878 deaths (34%):
  - **Public training set (70%):** data and outcomes available to participants
  - **Private test set (30%):** held out for evaluation
  - Electronic medical records (EMR) + pre-treatment CT imaging

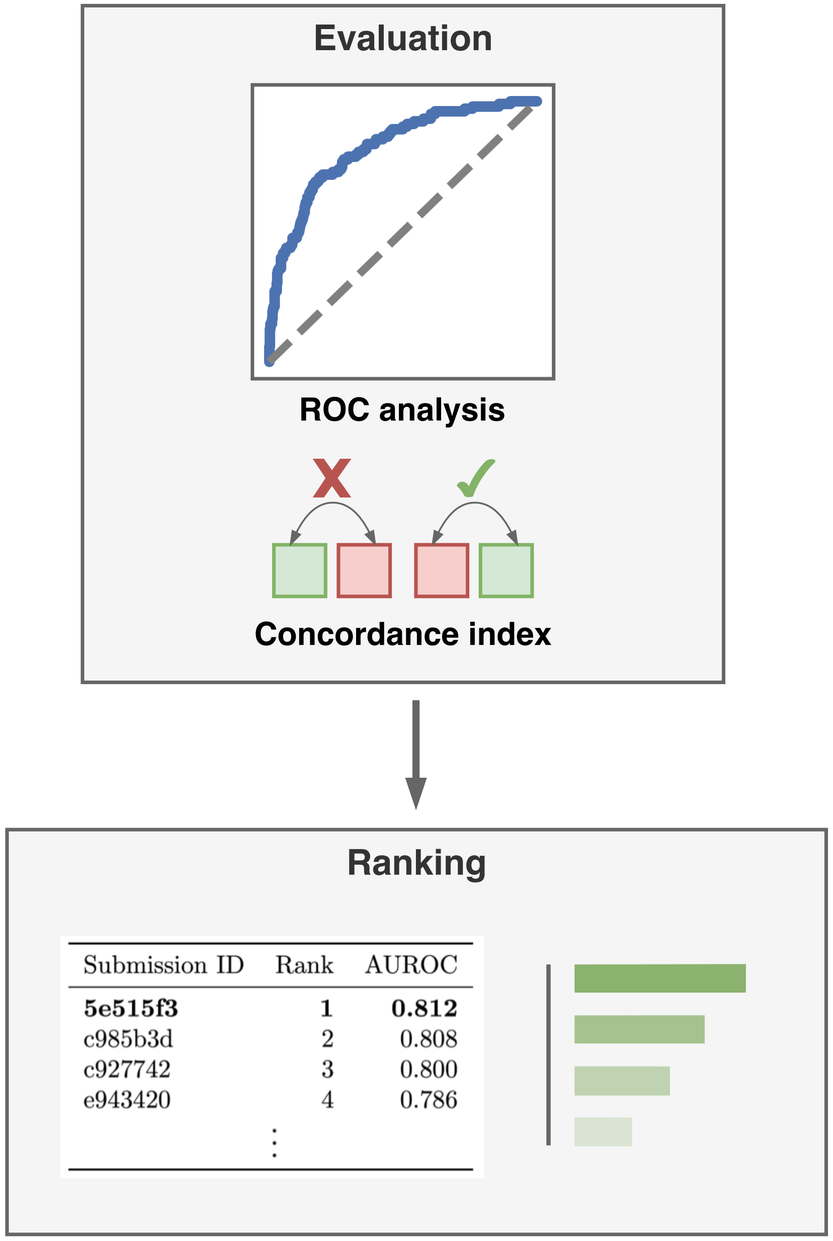
- Same training/test data for all participants ➛ can perform **unbiased comparison between methods**



Evaluation

- **Prediction endpoints**
  - 2-year survival (binary, main endpoint)
  - Lifetime risk of death (time-to-event)

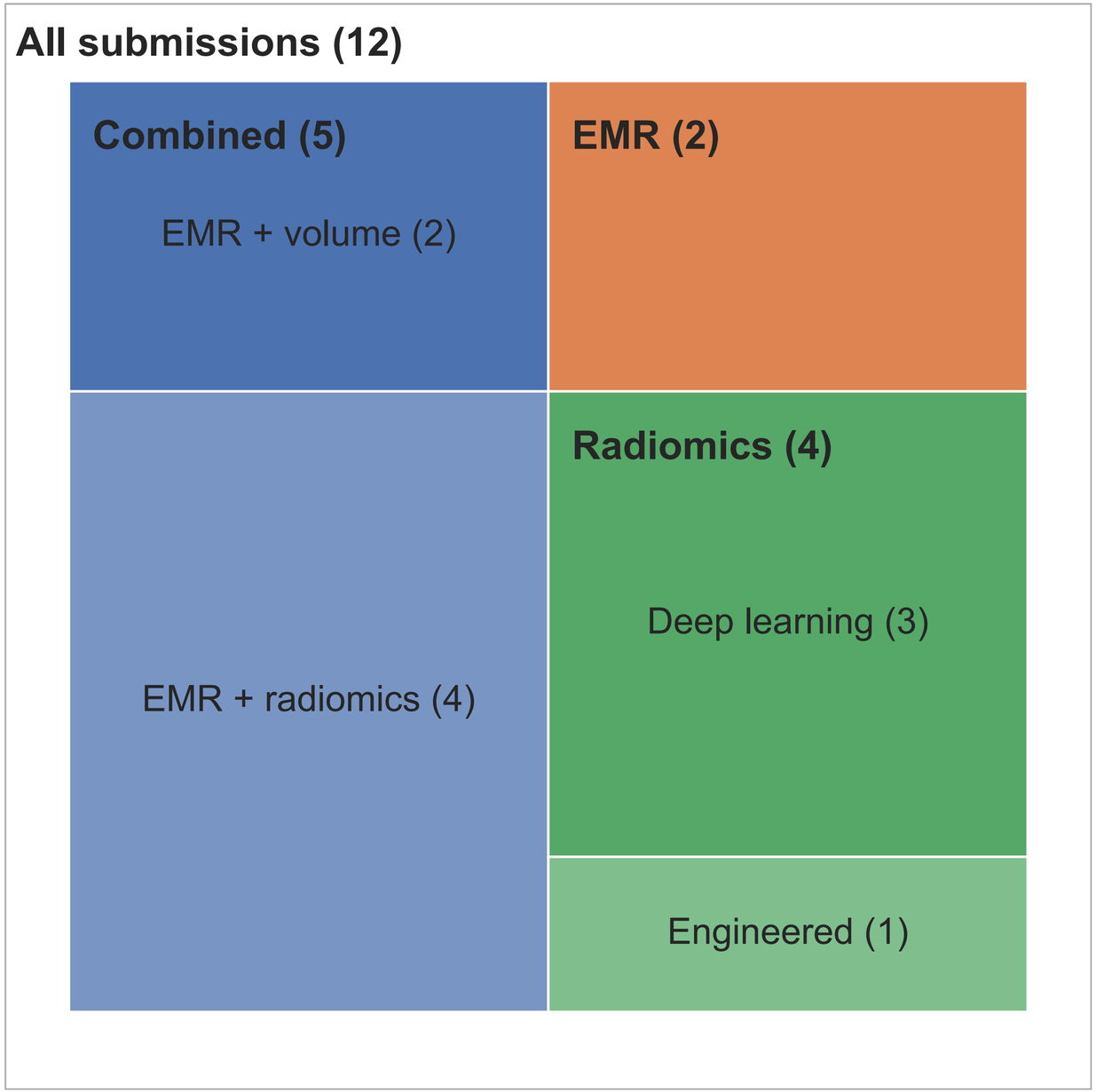
- **Evaluation metrics**
  - **Binary:** area under ROC curve (AUROC)
  - **Lifetime risk:** concordance (*C*-index)



Submission ID	Rank	AUROC
5e515f3	**1**	**0.812**
c985b3d	2	0.808
c927742	3	0.800
e943420	4	0.786
	⋮	

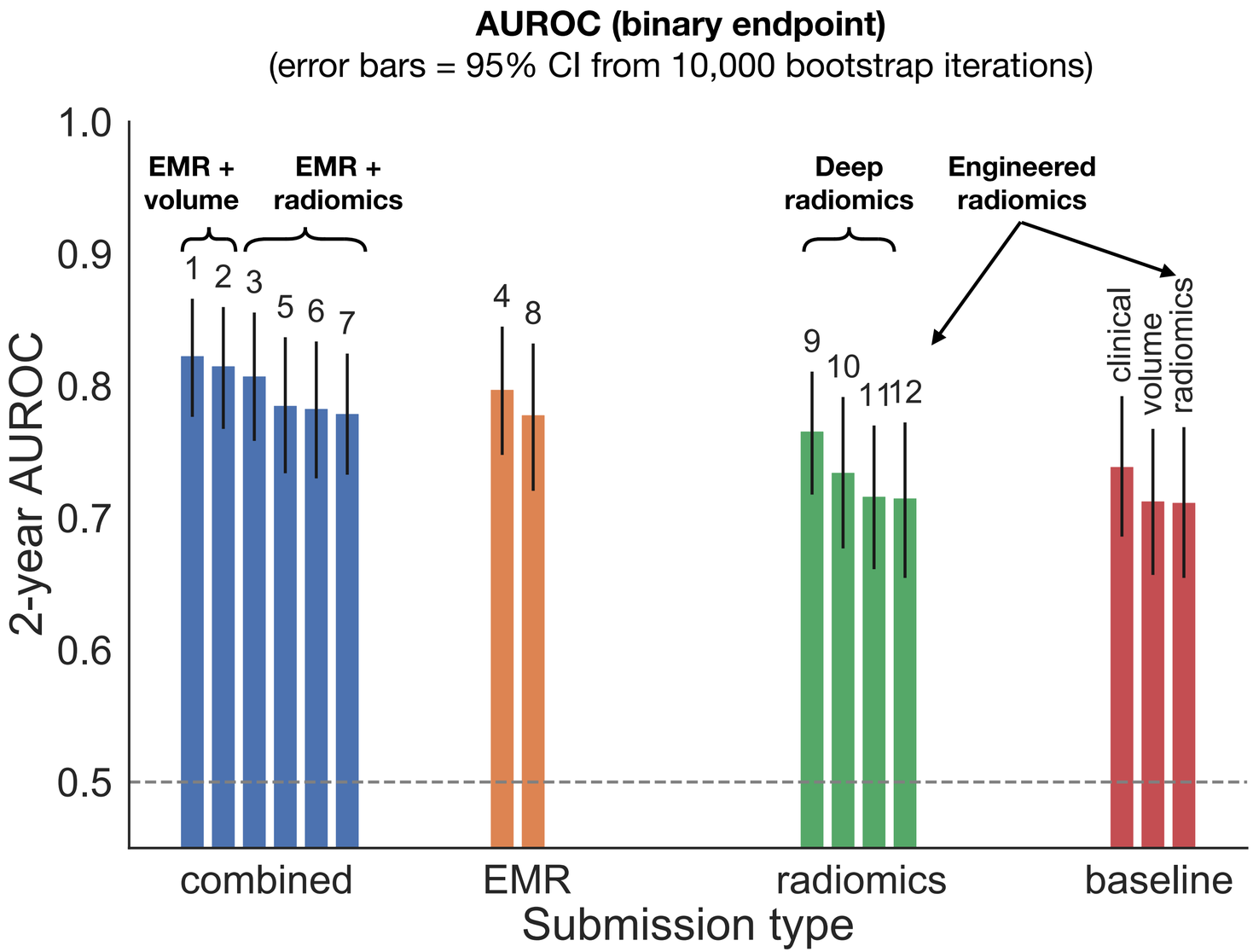
Overview of submissions

- **12 submissions overall**
- **3 major groups:**
  - **Combined:** using both EMR and imaging data
  - **EMR:** using EMR data only
  - **Radiomics:** using imaging data only



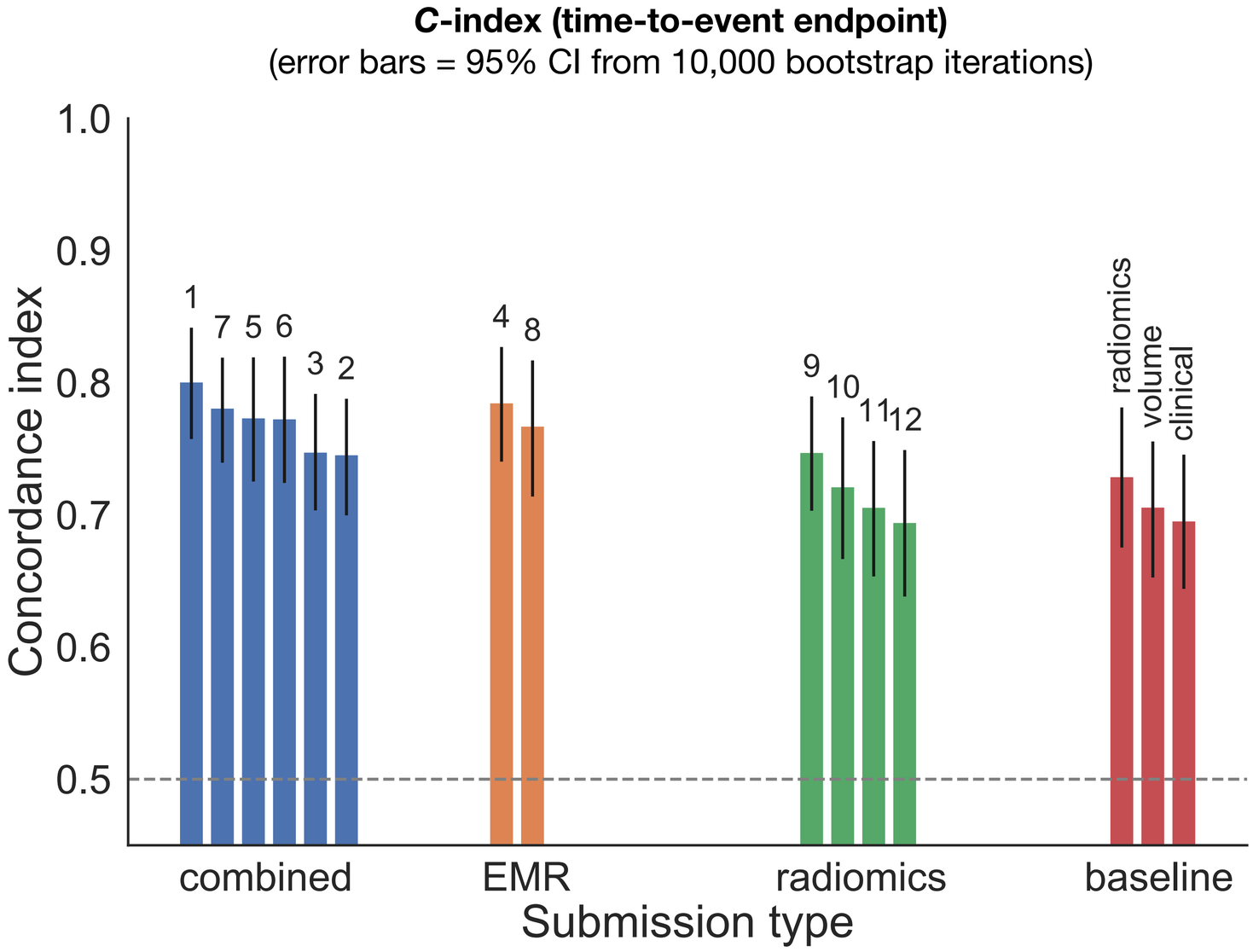
Challenge results

- **Deep learning radiomics** performs better than engineered on average
- **Small performance gap between EMR and combined** → EMR features drive predictions
- **Best combined submissions** use EMR features and tumour volume



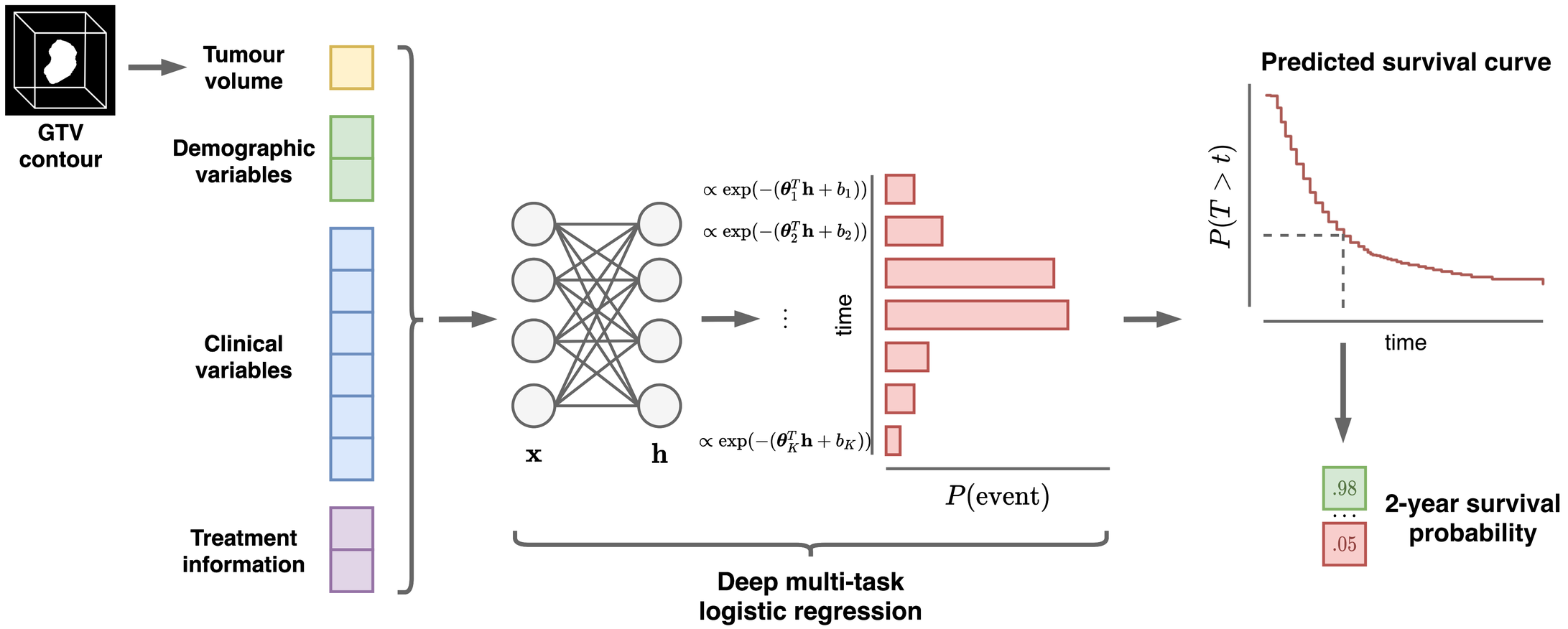
Challenge results

- **Deep learning radiomics** performs better than engineered on average
- **Small performance gap between EMR and combined** → EMR features drive predictions
- **Best combined submissions** use EMR features and tumour volume

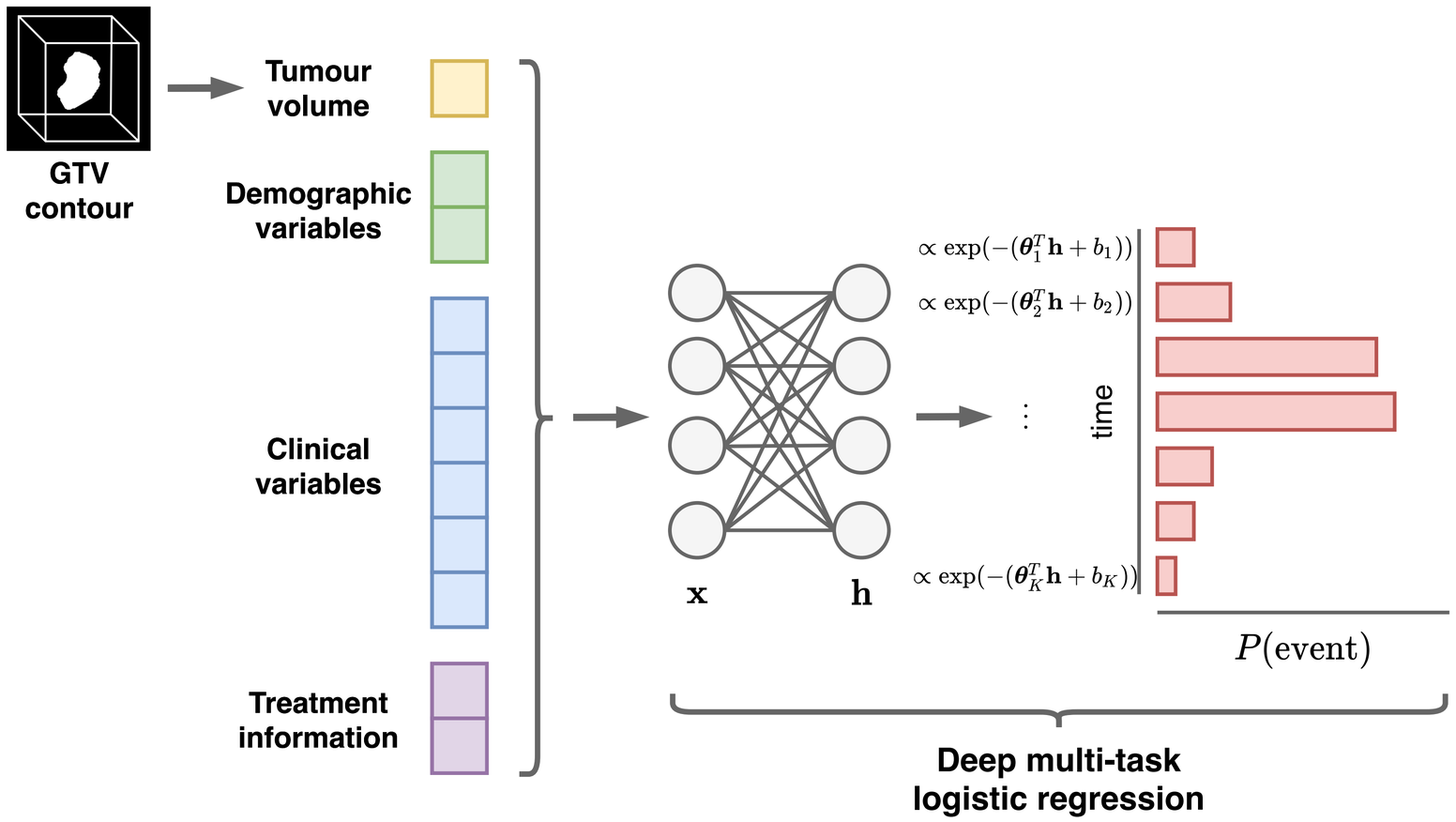


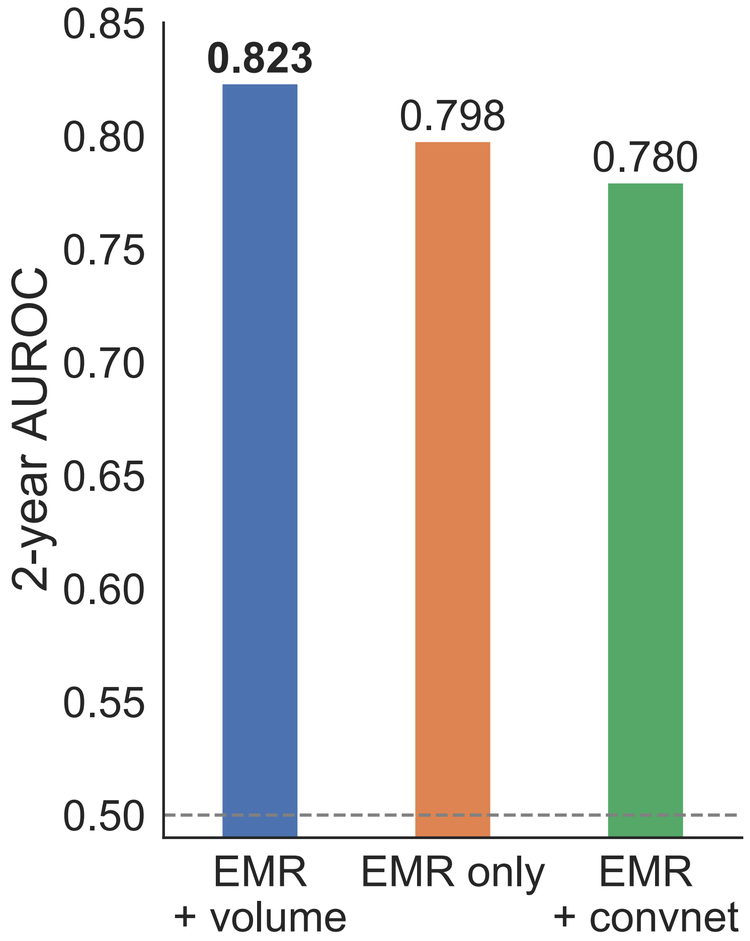
Good agreement between metrics (Pearson $r = 0.82$)

Best approach overall relies on multi-task learning and simple image feature



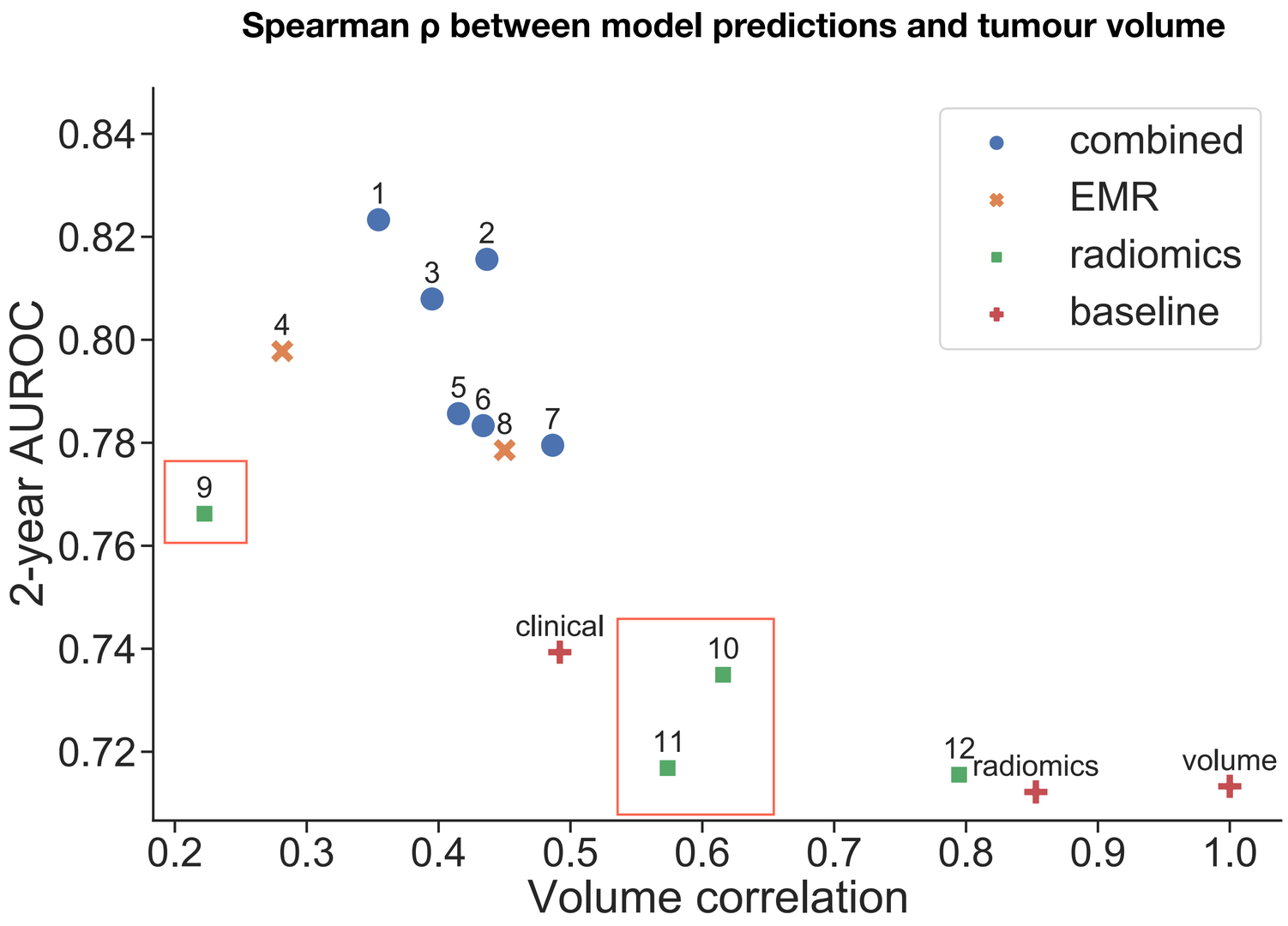
Best approach overall relies on multi-task learning and simple image feature





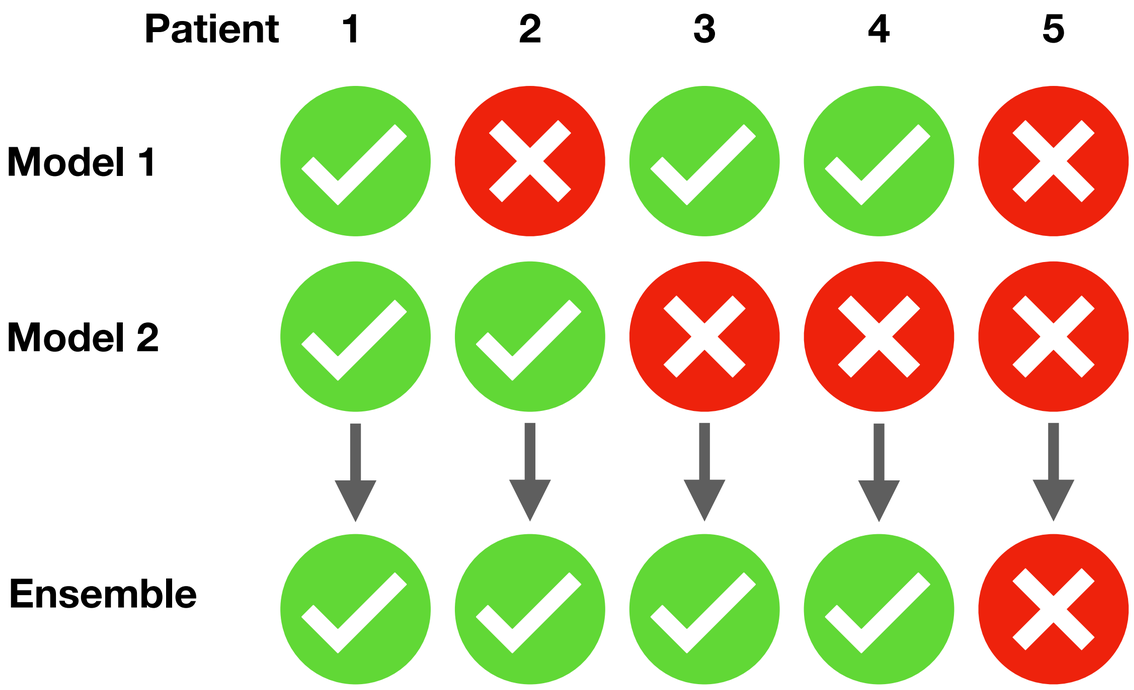
Correlation of predictions with tumour volume

- Best models show relatively low correlation while achieving good performance
- Most radiomics submissions show high correlation of predictions with volume ($\rho \geq .5$)
- Weaker correlation for deep learning models (9, 10, 11) than engineered features (12, `radiomics`)



Ensemble of all submissions achieves superior performance

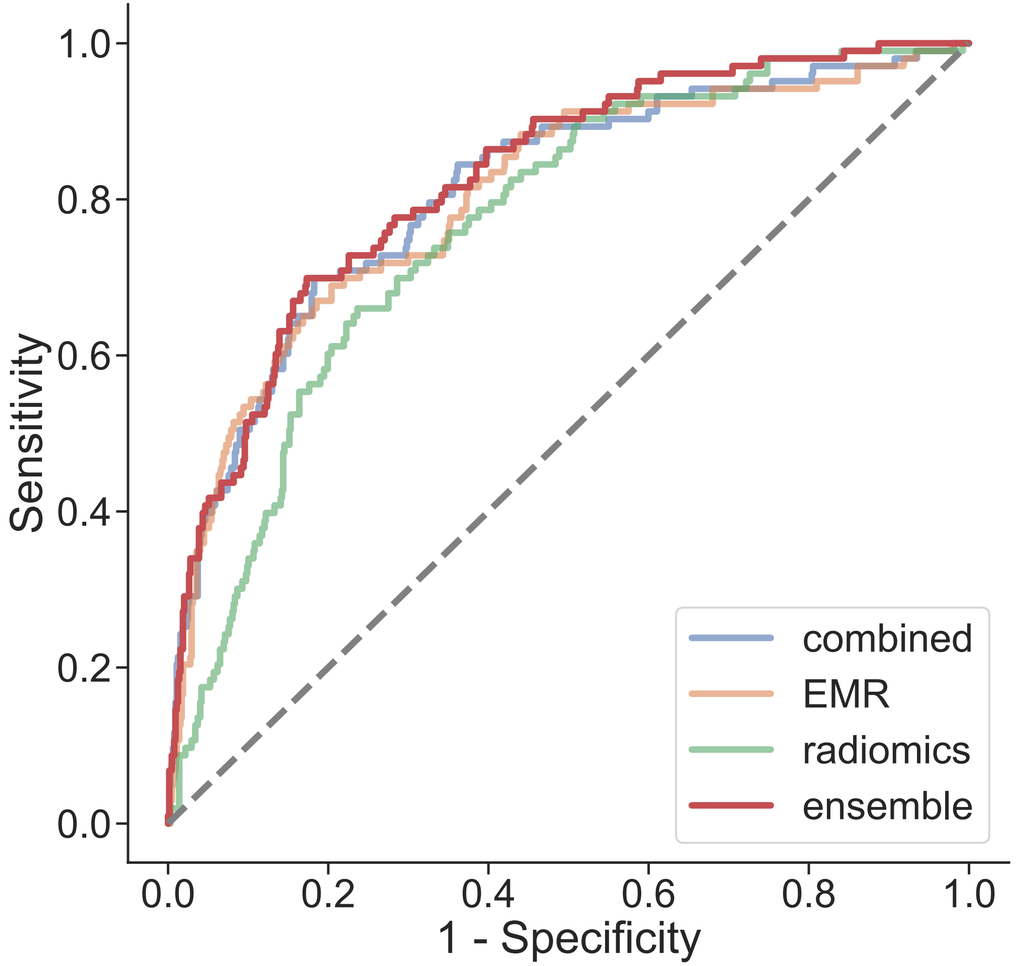
- **Ensembling** combines predictions of different models to improve performance ('wisdom of the crowds')



Ensemble of all submissions achieves superior performance

- **Ensembling** combines predictions of different models to improve performance ('wisdom of the crowds')

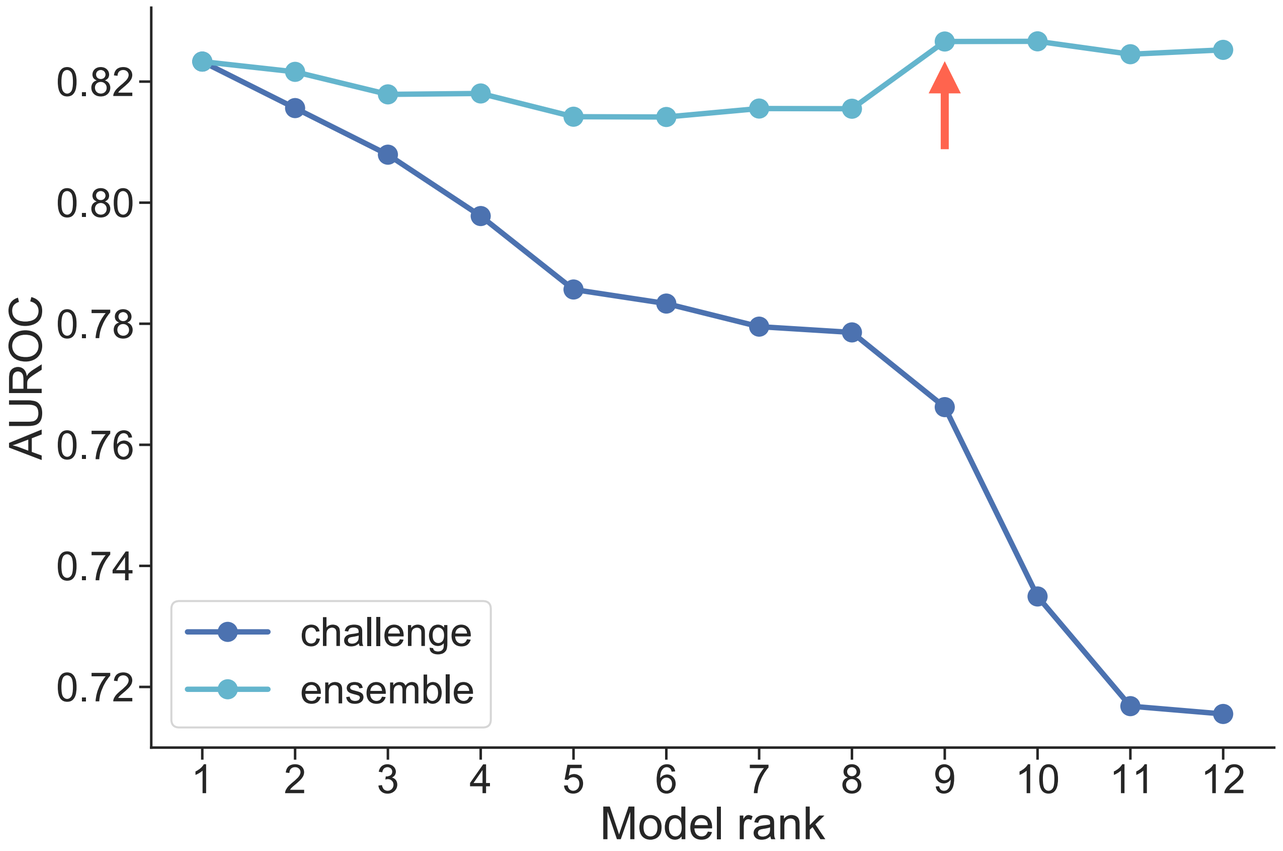
- Ensemble of all submissions slightly outperforms the best individual model



kind	AUROC	C-index
ensemble	**0.825 [0.781–0.866]**	**0.808 [0.770–0.843]**
combined	0.823 [0.777–0.866]	0.801 [0.757–0.842]
EMR	0.798 [0.748–0.845]	0.785 [0.740–0.827]
radiomics	0.766 [0.718–0.811]	0.748 [0.703–0.790]

Ensemble of all submissions achieves superior performance

- **Ensembling** combines predictions of different models to improve performance ('wisdom of the crowds')
- Ensemble of all submissions slightly outperforms the best individual model
- Largest performance improvement from the **best radiomics submission**



	1	2	3	4	5	6	7	8	9	10	11	12
best submission	1.00	0.94	0.94	0.92	0.82	0.76	0.77	0.78	0.55	0.51	0.46	0.35
volume	0.35	0.44	0.40	0.28	0.42	0.43	0.49	0.45	0.22	0.62	0.57	0.79

Correlation with individual predictions

Summary

- The RADCURE challenge enabled transparent and rigorous comparison of a diverse set of prognostic models in a large HNC dataset
- The winning submission used EMR features and volume together with a deep multi-task learning approach
- Deep learning on pre-treatment CT images achieved good performance (better than engineered radiomics and volume), but fell short of EMR + volume (even in combination with EMR features)
- Combining all submissions in an ensemble model yielded improved performance, and the biggest gain is from a deep radiomics model
- We plan to share the dataset and code with the upcoming publication

Acknowledgements

BHK lab

- **Dr. Benjamin Haibe-Kains**
- **Dr. Mattea Welch**
- Dr. Arvind Mer
- Dr. Wail Ba-Alawi
- Dr. Farnoosh Khodakarami
- Dr. Soheil Jahangiri-Tazehkand
- Dr. Sisira Nair
- Dr. Reza Reiazi
- Dr. Brad Van Oosten
- Dr. Jun Won Kim
- Petr Smirnov
- Ian Smith
- Alex Adam
- Joseph Marsilla
- **Sejin Kim**
- Colin Arrowsmith
- Anthony Mammoliti
- Chris Eeeles
- Minoru Nakano
- Gangesh Beri
- Nikta Feizi

Princess Margaret Cancer Centre

- Dr. Andrew Hope
- Zhibin Lu
- Dr. Tony Tadic
- Tirth Patel

Supervisory committee

- Dr. Scott Bratman
- **Dr. Chris McIntosh**

UHN Princess Margaret Cancer Centre

Medical Biophysics UNIVERSITY OF TORONTO

STARS 21 STRATEGIC TRAINING IN TRANSDISCIPLINARY RADIATION SCIENCE