

1 Foundation Models for Quantitative Biomarker Discovery in Cancer 2 Imaging

3
4 **Authors** | Suraj Pai^{1,2,3}, Dennis Bontempi^{1,2,3}, Vasco Prudente^{1,2,3}, Ibrahim Hadzic^{1,2,3}, Mateo Sokač^{4,5}, Tafadzwa
5 L. Chaunzwa^{1,3}, Simon Bernatz^{1,3}, Ahmed Hosny^{1,3}, Raymond H Mak^{1,2}, Nicolai J Birkbak^{4,5}, Hugo JWL Aerts^{1,2,3,6}
6

7 **Affiliations** | ¹ Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical
8 School, Harvard Institutes of Medicine, 77 Avenue Louis Pasteur, Boston, MA 02115, United States of America;
9 ²Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Universiteitssingel 40, 6229 ER
10 Maastricht, The Netherlands; ³Department of Radiation Oncology, Brigham and Women's Hospital, Dana-
11 Farber Cancer Institute, Harvard Medical School, 75 Francis Street and 450 Brookline Avenue, Boston, MA
12 02115, USA; ⁴Department of Molecular Medicine, Aarhus University Hospital, 8200 Aarhus, Denmark;
13 ⁵Department of Clinical Medicine, Aarhus University, 8200 Aarhus, Denmark; ⁶Department of Radiology,
14 Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, 75 Francis Street and
15 450 Brookline Avenue, Boston, MA 02115, USA;
16

17 **Running title** | Foundation Model for Cancer Imaging Biomarkers
18

19 **Corresponding author** | Hugo JWL Aerts, Ph.D., Artificial Intelligence in Medicine (AIM) Program, Mass
20 General Brigham, Harvard Medical School, Harvard Institutes of Medicine – HIM 343, 77 Avenue Louis Pasteur,
21 Boston, MA 02115, P - 617.525.7156, F - 617.582.6037, Email: Hugo_Aerts@DFCI.harvard.edu
22
23

Abstract | Foundation models represent a recent paradigm shift in deep learning, where a single large-scale model trained on vast amounts of data can serve as the foundation for various downstream tasks. Foundation models are generally trained using self-supervised learning and excel in reducing the demand for training samples in downstream applications. This is especially important in medicine, where large labeled datasets are often scarce. Here, we developed a foundation model for imaging biomarker discovery by training a convolutional encoder through self-supervised learning using a comprehensive dataset of 11,467 radiographic lesions. The foundation model was evaluated in distinct and clinically relevant applications of imaging-based biomarkers. We found that they facilitated better and more efficient learning of imaging biomarkers and yielded task-specific models that significantly outperformed their conventional supervised counterparts on downstream tasks. The performance gain was most prominent when training dataset sizes were very limited. Furthermore, foundation models were more stable to input and inter-reader variations and showed stronger associations with underlying biology. Our results demonstrate the tremendous potential of foundation models in discovering novel imaging biomarkers that may extend to other clinical use cases and can accelerate the widespread translation of imaging biomarkers into clinical settings.

INTRODUCTION

Foundation models present a paradigm shift in deep learning wherein a model trained on vast amounts of unannotated data can serve as the foundation of a wide range of downstream tasks. Recently foundation models have provided unprecedented performance gains in language, vision, and several other domains¹. In the field of natural language processing (NLP), for example, foundation models drive the successes of applications such as ChatGPT², BERT³, and CLIP⁴. Similarly, foundation models, such as SimCLR⁵ and DINO⁶, have reported considerable success in computer vision applications.

Medicine represents a vast potential for foundation models as labeled data are scarce, while multimodal data, such as medical images, biologic, and clinical notes, are frequently collected in routine clinical care⁷. Indeed, different applications of foundation models, such as augmented surgical procedures, bedside decision support, interactive radiology reports, and note-taking, have been reported⁸.

While many studies investigating imaging-based biomarkers incorporate supervised deep learning algorithms into their models⁹⁻¹¹, they are typically applied in scenarios where large datasets are available for training and testing. The quantity and quality of annotated data are strongly linked to the robustness of deep learning models. Access to large amounts of annotated data for specialized applications is often challenging and demands expertise, time, and labor. In such scenarios, many investigators fall back on traditional handcrafted or engineered approaches based on defined mathematical and statistical algorithms that analyze attributes like the shape and texture of objects in images, which limit the scope of discovery. This caveat is

59 commonplace in many scenarios where insights from imaging-based biomarkers have great potential in
60 informing clinical care.

61 Foundation models are generally pre-trained using self-supervised learning (SSL), a set of methods
62 that leverage innate information available within data by learning generalized, task-agnostic representations
63 (features) from large amounts of unannotated samples. Existing literature¹² has suggested several strategies
64 to pre-train networks to learn these representations. Approaches such as defining pre-text tasks that distort
65 an image and attempt to reconstruct the original or contrastively learning similar representations for
66 augmented views of the same image have primarily been investigated. Following pre-training, foundation
67 models can be applied to task-specific problems, improving generalization, especially in tasks with small
68 datasets. The expanding literature on SSL in medical imaging¹³ focuses primarily on two-dimensional images
69 (X-ray, whole slide images, dermatology images, fundus images, etc.) and diagnostic applications. There is still
70 limited evidence investigating whether SSL can help train foundation models that learn general, robust, and
71 transferrable representations that can act as imaging biomarkers, especially prognostic, for tasks of clinical
72 relevance.

73 In this study, we investigated whether foundation models pre-trained using self-supervised learning
74 can improve the development of deep learning-based imaging biomarkers. The foundation model was pre-
75 trained on 11,467 diverse and annotated lesions identified on computed tomography (CT) imaging from 2,312
76 unique patients¹⁴. The model was first technically validated on the classification of anatomical site lesions (use-
77 case 1). Subsequently, it was applied to two distinct clinically relevant applications: the development of a
78 diagnostic biomarker that predicts the malignancy of lung nodules (use-case 2) and a prognostic biomarker for
79 non-small cell lung cancer tumors in confirmed cancer cases (use-case 3). We evaluated two distinct
80 approaches of how a pre-trained foundation model can be incorporated into training pipelines for
81 downstream tasks, a direct approach of using the foundation model as a feature extractor combined with a
82 linear classifier and another approach where the foundation model is fine-tuned through deep learning. The
83 performance of the foundation model approaches was then evaluated and compared to conventional
84 supervised approaches in the three clinical use cases. Our analysis delves into limited data scenarios,
85 evaluating test-retest and inter-reader stability, determining explainability and interpretability through deep-
86 learning attribution methods, and exploring biological associations with gene expression data. Our results
87 demonstrate the potential of foundation models in discovering novel imaging biomarkers and their particular
88 strength in applications with limited datasets. This evidence may extend to other clinical use cases and imaging
89 modalities and can accelerate the widespread development and translation of imaging biomarkers into clinical
90 settings. To allow effortless incorporation, external evaluation, and validation, we are providing open access
91 to the foundation model along with reproducible workflows.

92

93

94

95 RESULTS

96 We developed a foundation deep learning model using SSL and tested the model's performance in three
97 distinct use cases. The study design and the pre-training process are outlined in **Fig. 1**. We developed the
98 foundation model using a dataset with 11,467 annotated CT lesions identified from 2,312 unique patients.
99 Lesion findings were diverse and included multiple lesions, such as lung nodules, cysts, and breast lesions,
100 among numerous others. A task-agnostic contrastive learning strategy was used to pre-train the model on
101 these lesion findings (see **Fig. 1a**), which subsequently was evaluated in three diverse clinical applications and
102 five distinct datasets (see **Fig. 1b**).

103
104 **Lesion anatomical site classification (Use-case 1)**. As a technical validation of the performance of the
105 foundation model, we selected an in-distribution task (i.e., sourced from the same cohort as that of the
106 foundation model pre-training) on 5,051 annotated lesions (see Use-case 1 in **Fig. 1b**). These specific lesions,
107 however, were not included in the pre-training data. Classification models were developed to predict the
108 correct anatomical site using a training and tuning dataset totaling 3,830 lesions. On an independent test set
109 of 1,221 lesions, we evaluated the performance of two different implementations of the foundation model
110 (see **Fig. 1c**).

111 We found that the foundation model approaches significantly outperformed the current standard
112 supervised approach using a randomly initialized model (i.e., random initialization of weights; see **Fig. 1d**) in
113 terms of balanced accuracy (BA) and mean average precision (mAP) (see **Fig. 2a, b**). When comparing
114 classification performances, the foundation features-based classifier (0.779 [95% CI 0.749-0.809], $p < 0.01$) and
115 the fine-tuned foundation model (0.804 [95% CI 0.773-0.834], $p < 0.01$), significantly improved BA ($p < 0.01$) over
116 the supervised model (0.72, [95% CI 0.689-0.750], $p < 0.01$) (see **Fig. 2a**). In terms of mAP, the fine-tuned
117 foundation model (0.856, [95% CI 0.828-0.886], $p < 0.01$) provided a significant ($p < 0.01$) performance benefit
118 over the supervised model (mAP=0.818 [95% CI 0.779-0.847], $p < 0.01$) (see **Fig. 2b**)

119 The performance advantage of the foundation model was even stronger in limited data scenarios (see
120 **Fig. 2a, b**). When we reduced training data to 50% ($n=2526$), 20% ($n=1010$), and 10% ($n=505$), the foundation
121 model as a feature extractor significantly improved BA and mAP over the supervised model. The fine-tuned
122 foundation model also significantly improved over the supervised model but failed to improve when training
123 data was reduced to 10%. Individual comparisons between each model at different data percentages can be
124 found in the supplementary material (see **Extended Data Table 1**).

125 To investigate feature separability, which indicates how well features can discriminate between
126 anatomical sites, we used dimensionality reduction methods to visualize features generated on the test set by
127 the foundation and the trained supervised models. The features from the foundation model produced
128 semantically separable clusters for each anatomical site, while features from the supervised model showed
129 poor separability (see **Fig. 2c-d**). Of note, unlike the supervised model, the foundation model was not exposed
130 to anatomical site information during training.

131

132 **Nodule malignancy prediction (Use case 2).** To assess the robustness of the foundation model, we chose an
133 out-of-distribution task (i.e., belonging to a different cohort than that of the foundation model training data)
134 involving predicting the malignancy of lung nodules from the LUNA16 dataset (see Use-case II in **Fig. 1b**). We
135 conducted our training on a labeled subset of 507 lung nodules with indications of malignancy suspicion. On
136 an independent test set of 170 nodules, we evaluated the performance of the two foundation model
137 implementations and two supervised learning approaches - random initialization and fine-tuning from another
138 supervised model. The model trained in use case 1 was chosen for the supervised fine-tuning.

139 The approach of fine-tuning the foundation model resulted in significant ($p < 0.01$) superiority over
140 both the supervised learning approaches (see **Fig. 3a, b**). The fine-tuned foundation model achieved an area-
141 under receiver operating curve (AUC) of 0.944 (95% CI 0.914-0.982, $p < 0.01$) and mAP of 0.952 (95% CI 0.926-
142 0.986, $p < 0.01$) compared to the fine-tuned supervised model's AUC of 0.857 (95% CI 0.806-0.918, $p < 0.01$) and
143 mAP of 0.874 (95% CI 0.822-0.936, $p < 0.01$).

144 When analyzing reduced data sizes, the fine-tuned foundation model significantly ($p < 0.01$)
145 outperformed the fine-tuned supervised model when data was reduced to 50% ($n=254$) and 20% ($n=101$).
146 However, it did not significantly improve when data was reduced to 10% ($n=51$). In contrast, the foundation
147 model as a feature-extractor improved significantly ($p < 0.005$) over all other models at 10%. Moreover,
148 performance from the foundation model as a feature extractor remained relatively stable even when trained
149 on 10% of the data, while all other models showed a significant drop in performance. Across the limited data
150 evaluation, although fine-tuned supervised models showed a trend of improvement over randomly initialized
151 supervised models, they were not found to be significant ($p > 0.05$). Detailed comparisons can be found in the
152 supplementary material (see **Extended Data Table 2**)

153 We observed that representations from the foundation model demonstrated superior linear
154 discrimination compared to the supervised model, where samples remained interspersed between the classes
155 (see **Fig. 3c, 3d**).

156

157 **Prognostication performance for non-small cell lung cancer (NSCLC) tumors (Use case 3).**

158 In the last use case, we evaluated the ability of the foundation model to capture quantitative radiographic
159 phenotypes of NSCLC tumors and consequently determine the prognosis of patients using three independent
160 cohorts of patients treated with surgery or radiation, HarvardRT ($n=291$), LUNG1 ($n=421$) and RADIO ($n=144$)
161 (see use-case 3 in **Fig. 1b**). We aimed to investigate the performance of foundation model implementations
162 when trained and applied to cohorts with strong distribution shifts (cohorts from separate institutions with
163 different standards of care). Therefore, we trained and tuned our prognostication models using data from the
164 HarvardRT cohort to predict 2-year overall survival after treatment and then compared the performance of
165 the foundation model and supervised approaches on the LUNG1 and RADIO cohorts.

166

167 In the LUNG1 cohort, foundation models outperformed both supervised methods, with statistical significance
168 ($p < 0.05$). Features extracted from the foundation model obtained an AUC of 0.637 (95% CI 0.583-0.691), and
169 fine-tuning the foundation model resulted in an AUC of 0.619 (95% CI 0.564-0.674), as shown in **Fig. 4a**. In
170 comparison, training supervised models with randomly initialized weights resulted in an AUC of 0.531 (95% CI
171 0.475-0.587). Fine-tuning a supervised model trained on a different task (use-case 1) showed an AUC of 0.566
172 (95% CI 0.510-0.622). The best-supervised model (supervised fine-tuned) and the foundation model (features
173 + linear classifier) were evaluated using Kaplan-Meier survival analysis, shown in **Fig. 4c** and **4e**, respectively.
174 The foundation model demonstrated higher prognostic power by better stratifying mortality, as shown by a
175 lower p-value ($p < 0.0001$) when split by the median on the tuning set, compared to the supervised model
176 ($p = 0.03$). Kaplan-Meier curves and univariate Cox regression for all of the models can be found in the
177 supplementary (see **Extended Data Fig. 1, Table 3**)

178 In the RADIO cohort, the foundation model as a feature extractor performed the best, with an AUC of
179 0.61 (95% CI 0.501-0.720). Supervised models trained with random initialization had an AUC of 0.532 (95% CI
180 0.426-0.639) while fine-tuning a supervised model led to an AUC of 0.567 (95% CI 0.468-0.665). Fine-tuning
181 the foundation model did not improve performance, yielding an AUC of 0.532 (95% CI 0.428-0.636), as shown
182 in **Fig. 4b**. Using foundation model features was significantly better than the randomly initialized supervised
183 model ($p < 0.05$), but none of the other networks showed significant differences from the rest ($p > 0.05$). Kaplan-
184 Meier survival analysis demonstrated significant stratification for the feature-extractor foundation model
185 predictions ($p = 0.008$) compared to the fine-tuned supervised model ($p = 0.138$), as shown in **Fig. 4d** and **4f**.
186 Kaplan-Meier curves and univariate Cox regression for all of the models can be found in the supplementary
187 material (see **Extended Data Fig. 1, Table 3**).

188
189 **Stability of the foundation model.** We evaluated the stability of our foundation model and compared it against
190 supervised approaches in two ways: through a test-retest scenario and an inter-reader variability analysis. To
191 assess test-retest robustness, we used scans from 26 patients from the RIDER dataset¹⁵ taken within a 15-
192 minute interval using the same imaging protocol. We found that predictions from the best-performing models,
193 feature-extractor foundation, and fine-tuned supervised had high stability with intraclass correlation
194 coefficient (ICC) values of 0.98 and 0.97, respectively. Furthermore, the test-retest features for both networks
195 were strongly correlated (as shown in **Extended Data Fig. 2a** and **2b**).

196 To evaluate stability against inter-reader variability, we used the LUNG1 dataset and perturbed the
197 input seed point to extract the 3D volume, simulating variations among human readers. We found that the
198 feature-extractor foundation models had higher stability against simulated inter-reader variations in
199 prediction performance than the fine-tuned supervised models (see **Extended Data Fig. 2c** and **2d**).

200
201 **Saliency maps for fine-tuned foundation models.** To gain insight into the regions of the input volumes that
202 contribute to a given prediction, we employed gradient-based saliency maps for foundation models fine-tuned
203 on three selected use cases (as depicted in **Fig. 5**). We used smooth guided back-propagation^{16,17} to compute

204 the gradient of the output with respect to the input while keeping the model weights constant. This provided
205 insight into the regions of the input that had the most significant influence on the output prediction.

206 Our analysis revealed that fine-tuned foundation models for each use case focused on different
207 regions but largely converged on tissues within or in proximity to the tumor. This is consistent with research
208 demonstrating the tumor microenvironment's influence on cancer development¹⁸ and prognosis. Specifically,
209 lesion anatomical site classification models (as depicted in **Fig. 5a**) focused mainly on areas surrounding the
210 lesions, such as the parenchyma and bone regions in the lung and the trachea in mediastinal lesions. On the
211 other hand, nodule malignancy models (as depicted in **Fig. 5b**) primarily concentrated on the tissues of the
212 nodule while avoiding high-density bone regions. In the case of prognosis networks (as depicted in **Fig. 5c**),
213 the model predictions were primarily attributed to areas surrounding the center of mass of the tumor, with
214 some contribution from high-density bone regions. Overall, these findings indicated that the areas that
215 contribute to the networks' predictions varied in accordance with the specific use case, with the tumor and
216 surrounding tissues playing a pivotal role.

217

218 **Underlying biological basis of the foundation model.** Finally, we investigated the biological basis of our
219 foundation model by analyzing gene expression data associated with model predictions for 130 subjects from
220 the RADIO dataset. To identify relevant genes, we selected the top 500 genes and performed a correlation
221 analysis, comparing the feature-extractor foundation and fine-tuned supervised model predictions with gene
222 expression profiles. We found that absolute correlation coefficients between gene expression profiles and
223 model predictions were significantly higher ($p=0.008$) for the foundation model, indicating a stronger
224 association with underlying tumor biology (see **Fig. 6a**).

225 Additionally, we examined the genes associated with these models through a gene set enrichment
226 analysis (genes with a correlation coefficient > 0.1). Our analysis revealed that foundation models showed a
227 pattern of enrichment of immune-associated pathways, including interferon signaling, interferon gamma
228 signaling, MHC class II antigen presentation, and PD-1 signaling. Conversely, while the supervised model did
229 show enrichment of individual pathways, no identifiable pattern was observed (see **Fig. 6b**).

230

231

232

233

234

235

236

237

238

239 DISCUSSION

240 In this study, we demonstrated that our foundation model trained using self-supervised learning, provided
241 robust quantitative biomarkers for predicting anatomical site, malignancy, and prognosis across three
242 different use cases in four cohorts. Several studies^{19–21} have demonstrated the efficacy of self-supervised
243 learning in medicine where only limited data might be available for training deep learning networks. Our
244 findings complement and extend this for identifying reliable imaging biomarkers for cancer-associated use
245 cases. We showed that our foundation model provided superior performance for anatomical lesion site and
246 malignancy prediction. Modeling using features extracted from the foundation model was the most robust
247 across tasks offering stable performance even when data sizes were considerably reduced to 51 samples (10%
248 of use-case 2). These features could also categorize data from these tasks into semantically separable clusters
249 corresponding strongly with target classes, although these features were learned independent of class
250 information. Using these features provided the best performance on small cohorts in predicting prognosis and
251 also demonstrated significant stratification of patients by their associated risk for each of the LUNG1 and
252 RADIO cohorts ($p < 0.01$). Additionally, predictions using the foundation model features were found to be highly
253 stable against inter-reader (standard deviation=0.004) and test-retest variations (ICC=0.98). Regarding the
254 interpretability of features, we observed that models focused on varying regions of the tumor and surrounding
255 tissue relevant to the associated use case. To gain insight into the underlying biological associations of these
256 features, RNA sequencing analysis combined with imaging data showed that these features correlated with
257 immune-associated pathways.

258 Studies for predicting endpoints, such as overall survival on small cohorts largely rely on statistical
259 feature extraction (engineered radiomics) and classical machine learning-based modeling. Precise three-
260 dimensional segmentations are required for extracting these statistical features from tumor volumes
261 increasing the annotation burden associated with these studies. Moreover, these statistical features are
262 affected by several confounders, such as inter-reader variability in segmentations²² and acquisition settings
263 of the scanners²³. Deep learning methods, in comparison, are robust to differences in acquisition and
264 segmentation variability and provide improved performance over statistical features¹⁰. However, they remain
265 restricted in their applicability in such low-data scenarios due to their dependency on large amounts of data
266 to provide robust performance. Training deep-learning models on small cohorts often lead to overfitting,
267 which diminishes performance when external data is introduced¹¹. Our foundation model approach has
268 several innovations: first, we developed a deep-learning system on a large corpus of 3D lesion images with
269 considerable diversity in their presentation. To our knowledge, our study is the first to pre-train a deep-
270 learning model using 11,467 3-dimensional lesion volumes. Second, we demonstrated that our pre-trained
271 model learned generalizable features and improved performance across three tasks and associated endpoints.
272 Our model also provided prognostic value when trained on small cohorts and applied to external validation
273 cohorts. Third, our models showed high robustness to test-retest and inter-reader variations. Finally, we share

274 our validated foundation model with the public, allowing external testing and future studies to facilitate their
275 adoption into external workflows.

276 Several studies have investigated deep learning algorithms for identifying cancer imaging biomarkers
277 in both small and large cohorts. Hosny et al.¹⁰ trained a deep learning model for lung cancer prognostication
278 using several multi-institutional cohorts and demonstrated strong performance using deep learning methods
279 over traditional radiomics features. Kumar et al.²⁴ identified radiomic sequences using deep convolutional
280 encoders for determining the malignancy of lung nodules from the LIDC-IDRI dataset considering 4306 lesions.
281 Lao et al.²⁵ proposed a deep-learning model-based radiomics signature for predicting survival in glioblastoma
282 multiforme, trained and validated on relatively small cohorts. Haarbarger et al.²⁶ present a deep convolutional
283 network-based approach to predict survival endpoints on the LUNG1 dataset. Cho et al.²⁷ developed a
284 radiomics-guided deep-learning model for stratifying the prognosis of lung adenocarcinoma and validated it
285 in a local cohort and an external validation cohort. A general trend observed across these studies is that the
286 performance of deep learning models is more robust when larger and multi-institutional cohorts are available
287 for training. Validation is subsequently performed on cohorts smaller than the training cohort. A demonstrated
288 strength of our approach is that training on smaller cohorts performs well in larger validation cohorts. For the
289 prognostication use case, we performed well on two external validation cohorts with a combined size
290 considerably larger than the training cohort. Our pre-trained foundation model shows strong generalization
291 ability across our diverse use cases and may apply to several other cancer imaging use cases out of the box.
292 Furthermore, extracting features from our model (inference only) followed by simple modeling methods is
293 resource-efficient, alleviating the need for expensive hardware for training standard deep-learning models
294 while providing on-par performance.

295 In recent years, self-supervised pre-training has been applied to medical imaging with promising
296 results^{19,21,28,29}. Zhou et al.³⁰ present an approach that constructs several pre-text tasks to train SSL networks
297 and show that they outperform solely supervised networks trained across five clinically relevant tasks. A novel
298 contrastive SSL strategy incorporating both global and local information captured within medical images and
299 reporting their superior performance, especially in low-data settings, is proposed by Chaitanya et al.³¹. Azizi et
300 al.¹⁹ demonstrate that grouping multiple images attributed to the same medical condition along with
301 combining natural and medical images for contrastive SSL training improves performance. Specifically for deep
302 radiomics applications, Li et al.³² propose targeting data imbalance in existing data and present a combined
303 approach of traditional radiomic features and self-supervised learning representations, improving
304 performance for discriminating tumor grade and tumor staging tasks. Li et al.³³ proposed a novel self-
305 supervised collaborative approach for creating latent representations from radiomic features. Zhao and Yang³⁴
306 used self-supervised learning to pre-train models via a radiomic-deep feature correspondence task. Although
307 these studies have investigated self-supervised learning for radiomics tasks, they lacked external validation or
308 proposed limited evaluation of the generalizability of their approaches. Our study presents a foundation model

309 for radiomic discovery by pre-training on a large cohort of lesions. The examined tasks are independent of the
310 pre-training cohort and demonstrate the increased generalizability of our proposed approach.

311 Despite the strengths outlined in our study, we recognize several limitations that need to be addressed
312 prior to the clinical applicability of our foundation model. Features from the foundation model followed by
313 linear classifiers provided the most robust performance across all investigated tasks. However, linear classifiers
314 might be sub-optimal in identifying complex relationships between feature representations to predict
315 challenging endpoints. As we aimed to demonstrate the benefits of our foundation model compared to
316 existing approaches, we have limited our exploration with fine-grained feature and model selection strategies.
317 Comprehensive selection approaches similar to Parmar et al.³⁵ might improve performance even further,
318 strengthening our hypothesis for foundation models.

319 Similarly, deep learning-based finetuning approaches employed in this study are representative of
320 baseline performance. We observed that finetuning approaches for the foundation model in low data settings
321 (especially 10%) and smaller cohorts (HarvardRT) resulted in suboptimal performance compared to using
322 extracted features. We hypothesize that in lower data settings, models overfit the training data and
323 demonstrate worse generalization as the number of parameters to tune increases. However, with the steady
324 emergence of deep learning literature proposing improvements to handle aspects such as data imbalance,
325 hyperparameter selection, and optimization objectives, the performance of these models can be pushed far
326 above the current baseline. Our prognostication model is also limited in its performance due to our focus on
327 solely imaging data; incorporating clinical features has a large potential to improve its effectiveness.

328 Our foundation model's clinical applicability encounters challenges typically associated with deep
329 learning, including generalizability, interpretability, and explainability. Given the retrospective nature of this
330 study, our capacity to evaluate the real-world practicality of foundation model-based biomarkers is
331 constrained. Deep learning models are notorious for being black boxes that offer little clarity on interpretable
332 and explainable reasoning behind their predictions. Although we used well-established saliency attribution
333 methods to interpret our foundation model's predictions, the broader applicability of these insights is
334 hindered by the technical limitations of such methods^{36,37}. In addition to the limitations of deep learning
335 methodology, the biological association analysis conducted to explain our model's predictions is preliminary
336 and requires further investigation to generate a concrete understanding. We anticipate that future external
337 validation of our open-access model will help confront these prevalent challenges.

338 In conclusion, our foundation model offers a powerful and reliable framework for discovering cancer
339 imaging biomarkers, even in small datasets. Furthermore, it surpasses current deep learning techniques in
340 various tasks while fitting conveniently into existing radiomic research methods. This approach can potentially
341 uncover new biomarkers that significantly contribute to research and medical practice. We share our
342 foundation model and reproducible workflows so that more studies can investigate our methods, determine
343 their generalizability, and incorporate them into their research studies.

344 METHODS

345 **Study Population.** We utilize a total of five distinct datasets, four of which are publicly accessible, and one is
346 an internal dataset. These were acquired from various institutions as components of separate investigations
347 (see **Extended Data Table 4**).

348 DeepLesion¹⁴ is a dataset comprising 32,735 lesions from 10,594 studies of 4,427 unique patients
349 collected over two decades from the National Institute of Health Clinical Center PACS server. Various lesions,
350 including kidney, bone, and liver lesions - as well as enlarged lymph nodes and lung nodules, are annotated.
351 The lesions are identified through radiologist-bookmarked RECIST diameters across 32,120 CT slices. In our
352 study, we excluded CT scans with a slice thickness exceeding 3mm, resulting in 16,518 remaining lesions.
353 Subsequently, we divided this into 11,467 unlabelled lesions for contrastive training and 5,051 labeled lesions
354 for anatomical site classification. The labeled lesion data were further separated randomly into training,
355 tuning, and testing sets, containing 2,610, 1,220, and 1,221 lesions, respectively.

356 LUNA16³⁸ is a curated version of the LIDC-IDRI dataset of 888 diagnostic and lung cancer screening
357 thoracic CT scans obtained from seven academic centers and eight medical imaging companies comprising
358 1,186 nodules. The nodules are accompanied by annotations agreed upon by at least 3 out of 4 radiologists.
359 Alongside nodule location annotations, radiologists also noted various observed attributes like internal
360 composition, calcification, malignancy, suspiciousness, and more. For our evaluation, we chose nodules with
361 at least one indication of malignancy suspicion, totaling 677. We randomly picked 338 nodules for training and
362 169 for tuning the malignancy prediction networks. The final 170 nodules were utilized to assess the networks'
363 performance.

364 HarvardRT¹⁰ is a cohort of 317 patients with stage I-IIIB NSCLC treated with radiation therapy at the
365 Dana-Farber Cancer Institute and Brigham and Women's Hospital, Boston, MA, US, between 2001 and 2015.
366 All CT scans for this cohort were acquired with and without intravenous contrast on the GE Lightspeed CT
367 scanner. The primary tumor site was contoured by radiation oncologists using soft tissue and lung windows.
368 A subset of 291 patients with a follow-up of 2 years was selected for this study. We used 203 tumor volumes
369 for training the prognostication networks and the remaining 88 tumor volumes for tuning.

370 LUNG1³⁹ is a cohort of 422 patients with stage I-IIIB NSCLC treated with radiation therapy at MAASTRO
371 Clinic, Maastricht, The Netherlands. FDG PET-CT scans were acquired with or without contrast on the Siemens
372 Biograph Scanner. Radiation oncologists used PET and CT images to delineate the gross tumor volume. For our
373 study, we selected CT scans of 421 patients with annotated primary gross tumor volumes and used these as
374 an independent test set for prognostication networks.

375 RADIO (NSCLC-Radiogenomics)⁴⁰ dataset is a collection of 211 NSCLC stage I-IV patients recruited
376 between 2008 and 2012 who were referred for surgical treatment and underwent preoperative CT and PET/CT
377 scans. These patients were recruited from the Stanford University School of Medicine and the Palo Alto
378 Veterans Affairs Healthcare System. Scan scans were obtained using various scanners and protocols depending
379 on the institution and physician. A subset of 144 patients in the cohort has available tumor segmentations

380 independently reviewed by two thoracic radiologists. In addition to imaging data, the dataset includes
381 molecular data from EGFR, KRAS, ALK mutational testing, gene expression microarrays, and RNA sequencing.
382 For the current study, we utilized the subset of 144 patients with annotated gross tumor volumes as an
383 independent test set for prognostication and also investigated the biological basis of our networks using this
384 dataset.

385

386 **Data Preprocessing.** CT scans were resampled using linear interpolation to achieve isotropic voxels with a
387 1mm^3 resolution to address variations in slice-thickness and in-plane resolutions across study populations. We
388 extracted patches of $50 \times 50 \times 50$ voxels from the scans centered around a seed point (refer to **Extended Data**
389 **Fig. 3**). For the DeepLesion dataset, which provided annotations in the form of RECIST diameters, the seed
390 point was determined by calculating the midpoint of the RECIST diameter. For the other datasets (i.e., LUNA16,
391 HarvardRT, LUNG1, and RADIO), which supplied annotations as 3D contours, the seed point was obtained by
392 computing the center of mass (CoM). This approach allows for significantly higher throughput than manual
393 segmentation, which can be more tedious. We then normalized the voxel values in the patches by subtracting
394 -1024 (lower-bound Hounsfield unit) and dividing by 3072 (upper-bound Hounsfield unit), ensuring the
395 intensity values in the input data ranged between 0 and 1.

396

397 **Task-agnostic contrastive pre-training of the foundation model.** We implemented contrastive pre-training
398 using a modified version of the SimCLR framework⁵. The SimCLR framework's general principle involves
399 transforming a single data piece (e.g., a patch taken from a CT scan) into two correlated and augmented
400 samples (e.g., the same patch rotated 15 degrees clockwise and flipped horizontally). A convolutional encoder
401 is then used to extract latent representations from these samples. Through a contrastive loss function⁴¹, the
402 model learns to identify similar representations from the same data sample and dissimilar representations
403 from different data samples. The framework emphasizes effective transformation choices, convolutional
404 encoder architectures, and contrastive loss functions for optimal self-supervised learning performance. To
405 effectively represent the nature of medical images, we made modifications to each of these components.

406 Transformations proposed in the original SimCLR framework for natural world images, such as cutout
407 augmentation, Sobel filtering, and color distortion, are unsuited for 3D medical images due to dynamic range
408 and color depth differences. Therefore, our study applies different augmentations to replace these
409 transformations. For instance, we substituted the random color jitter transform with a random histogram
410 intensity shift transform, as they both induce variation in intensity distribution.

411 To extract representations from the transformed 3D volumes, we selected the 3D ResNet50
412 architecture as our deep convolutional encoder. While the SimCLR authors employed a 2D ResNet50
413 architecture, we opted for its 3D counterpart, which has proven effective in handling 3D medical imaging
414 data⁴².

415 Regarding loss functions, we extended normalized temperature-scaled cross-entropy loss (NT-Xent)⁴³
416 to support contrastive training for lesion volumes. The modifications include: 1) selecting positive pairs as 3D
417 patches surrounding the lesion's seed point, 2) choosing negative pairs by randomly sampling 3D patches from
418 the rest of the scan, and 3) computing the contrastive loss on these positive and negative pairs, with each
419 iteration comprising N positive pairs and $N*2(N-1)$ negative pairs. We also explored different temperature
420 parameters for the NT-Xent loss. However, the original value of 0.1 proposed by the original paper was the
421 most effective.

422 Our model was pre-trained for 100 epochs using an effective batch size of 64 (32 x 2 training nodes)
423 on two NVIDIA Quadro RTX 8000 GPUs taking approximately five days. We used Stochastic Gradient Descent
424 (SGD) as the optimizer, with layer-wise adaptive rate control (LARC), momentum, and weight-decay enabled.
425 To improve the optimization process, we employed learning rate schedulers that combined linear and cosine
426 decay strategies and a warmup phase to modify the learning rate at the beginning of training gradually. While
427 most specifications were consistent with the original SimCLR experiments, we experimented with different
428 batch sizes, patch sizes (50mm³ and 64mm³), learning rates, transforms, and model architectures.

429
430 **Task-specific training of the foundation model.** Our foundation model was adapted for a specific task through
431 two approaches: 1) extracting features and fitting a linear classifier on top of them or 2) fine-tuning the pre-
432 trained ResNet50 for the given classification task.

433 We extracted 4096 features from the foundation model for each data point and used them to train a
434 logistic regression model using the scikit-learn framework⁴⁴. A comprehensive parameter search for the
435 logistic regression model was performed using the optuna hyper-parameter optimization framework⁴⁵. No
436 performance improvements were observed through feature selection strategies; therefore, all 4096 features
437 were used in accordance with linear evaluation strategies prevalent in self-supervised learning (SSL) literature.

438 Fine-tuning was carried out with all layers updated during training, utilizing cross-entropy loss. A series
439 of randomly chosen augmentations—random flips, random 90-degree rotations, and random translations of
440 ± 10 voxels across all axes—were applied throughout the training. SGD was employed for network training,
441 with momentum enabled and step-wise learning rate decay. Following the original SimCLR experiments,
442 configurations and similar parameters (including learning rate, transforms, and model architectures) were
443 explored during hyperparameter tuning. Each network was trained for 100 epochs using a single NVIDIA
444 Quadro RTX 8000 GPU, and the best-performing model checkpoints was chosen based on the tuning set.

445 For supervised baseline models, their weights were initialized randomly, and they were trained using
446 the same configuration that was adopted for fine-tuning the foundation model. The supervised models for use
447 cases 2 and 3 were also fine-tuned, utilizing the same configuration as in the pre-trained fine-tuning process
448 but by initializing them with the weights of the trained supervised baseline from use case 1.

449 Task-specific training was conducted on reduced dataset sizes in addition to utilizing the entire
450 dataset. We randomly sampled 50%, 20%, and 10% of the training and tuning datasets and constructed task-

451 specific models using these samples with the same configuration as the entire dataset. As the training dataset
452 sizes decreased, we considered training the models for a higher number of epochs; however, models
453 frequently overfitted during extended training. The entire test dataset was employed to allow benchmarking
454 across these splits.

455
456 **Performance Analysis.** Validation of the foundation model was performed using several use-case-relevant
457 metrics. Lesion anatomical site classification performance was assessed using balanced accuracy (BA) as a
458 multi-label counting metric and mean average precision (mAP) as a multi-threshold metric. The multi-label
459 metric, BA, adjusts class-wise accuracy based on the class distribution at a chosen threshold (0.5). The multi-
460 threshold metric, mAP, enables the examination of a given class's performance across a range of prediction
461 thresholds. All classes other than the class of interest are considered negatives, and performance is averaged
462 across all possible classes. We avoided using the area under the receiver operating curve (AUC-ROC) for this
463 use case due to the high proportion of negatives relative to positives, which results in consistently low false-
464 positive rates and might overestimate the AUC. However, due to a more balanced class distribution, nodule
465 malignancy prediction was evaluated using AUC-ROC. NSCLC prognostication networks also employed AUC-
466 ROC for evaluation, as it estimates the ranking of subjects based on their survival times.

467 Models underwent pairwise comparison using permutation tests. N permutations (N=1000) were
468 conducted for each pair, and new models were computed after permuting class labels. Metrics were
469 recalculated after resampling, and a two-sided p-value was calculated to test the null hypothesis of
470 observations from each pair originating from the same underlying distribution. Additionally, 95% confidence
471 intervals were established for each model using a bootstrap test with N=9999 resamples.

472 Kaplan-Meier (KM) curves were also used to determine the stratification of subjects based on their
473 prediction scores for the prognostication models. Groups were selected based on prediction scores on the
474 tuning set, and curves were plotted on the test set for these groups. Multivariate log-rank tests were used to
475 examine the significance of the stratification. Univariate Cox regression models were built using the model
476 predictions as the categorical variables of interest, grouped similarly to the KM curve.

477
478 **Feature visualization and saliency maps.** We used the foundation and top-performing supervised models as
479 feature extractors to obtain 4096 distinct features per data point. To enable visual interpretation of these
480 high-dimensional features, we utilized t-SNE⁴⁶ (t-Stochastic Neighbourhood Embeddings) to reduce their
481 dimensionality to 2D. To arrive at the most interpretable visualization, we explored various parameter
482 configurations, including perplexity, initialization, and learning rates. Points in the 2D visualization were color-
483 coded according to their respective target classes, despite dimensionality reduction being agnostic to these
484 distinctions. Density contours were superimposed over the visualizations to enhance the understanding of
485 group patterns, offering a more comprehensive representation of trends across data points.

486 In order to generate saliency maps for each task, the fine-tuned foundation model was used to
487 generate predictions on randomly selected volumes from respective datasets. The fine-tuned foundation
488 model with a single output prediction (corresponding to the predicted target class) was chosen in contrast to
489 the feature extractor as computing saliency maps over 4096-dimensional outputs remains challenging in
490 practice. We used a combination of 1) smooth gradient backpropagation, which averages gradients of the
491 output with respect to several noisy inputs, and 2) guided back-propagation which combines deconvolution
492 with backpropagation, mainly stopping the flow of negative gradients or neurons that decrease the activation
493 signal. The method is termed smooth guided-backpropagation and is implemented in the MONAI framework
494 ⁴⁷.

495
496 **Stability Testing.** To test the stability of our models, we performed a test-retest stability and inter-reader
497 variation evaluation. For the test-retest evaluation, we compared model predictions (of outcome) from the
498 best foundation and supervised models generated on chest CT scans taken in a 15-minute interval for 32
499 patients. Intraclass correlation coefficient (ICC) was computed using the interrater reliability and agreement
500 package (*irr*) in R⁴⁸. We also tested the stability of the flattened features computed by the models by
501 calculating Spearman correlation and R².

502 For the inter-reader variation evaluation, we used the LUNG1 dataset and generated 50 random
503 perturbations sampled from a three-dimensional multivariate normal distribution with zero mean and
504 diagonal covariance matrix for each seed point. Across each dimension, a variance of 16 voxels was used for
505 generating samples. We generated predictions on perturbed seed points using the best foundation and
506 supervised model, resulting in 50 different prediction models for each. The mean and variance of the 50
507 models were computed for each and compared.

508
509 **Biological Associations.** The GSE103584 dataset contains 130 NSCLC (Non-Small Cell Lung Cancer) samples
510 that consist of paired CT scans and gene expression profiles generated by RNA sequencing. To analyze gene
511 expression profiles, we filtered them based on cohort mean expression and standard deviation. First, we took
512 only the genes with a higher expression than the overall dataset mean and then picked the top 500 genes
513 based on standard deviation. Next, we performed a correlation analysis comparing the best-supervised and
514 foundation models. To further evaluate foundation model features' association with tumor biology, we
515 computed the absolute value of the correlation coefficients and performed a gene set enrichment analysis
516 with all genes with a correlation coefficient above 0.1.

517
518
519
520
521
522

523 **ACKNOWLEDGEMENTS**

524 The authors acknowledge financial support from NIH (H.J.W.L.A: NIH-USA U24CA194354, NIH-USA
525 U01CA190234, NIH-USA U01CA209414, and NIH-USA R35CA22052), and the European Union - European
526 Research Council (H.J.W.L.A: 866504).

527

528 **AUTHOR CONTRIBUTIONS**

529 Study conceptualization: S.P, H.J.W.L.A.; Data acquisition, analysis, and interpretation: S.P, D.B, A.H, T.L.C,
530 H.J.W.L.A.; Methodological design and implementation: S.P, D.B.; Conceptualization of assessment strategies:
531 S.P, D.B, N.J.B, H.J.W.L.A; Statistical Analyses: S.P, M.S, N.J.B, H.J.W.L.A; Code and reproducibility: S.P, I.H, V.P;
532 Writing of the manuscript: S.P, D.B, M.S, S.B, H.J.W.L.A; Critical revision of the manuscript: All authors; Study
533 supervision: H.J.W.L.A

534

535 **DATA AVAILABILITY STATEMENT**

536 The majority of the datasets utilized in this study are openly accessible for both training and validation
537 purposes and can be obtained from the following sources: i) DeepLesion [nihcc.app.box.com/v/DeepLesion],
538 used both for our pre-training and use-case 1 ii) LUNA16 [luna16.grand-challenge.org] used for developing
539 our diagnostic image biomarker iii) LUNG1 [wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics]
540 and iv) RADIO [wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics] used for the validation
541 of our prognostic image biomarker model. The training dataset for our prognostic biomarker model,
542 HarvardRT, is internal and unavailable to the public. HarvardRT was collected under an IRB-approved
543 retrospective protocol with a waiver of consent (Dana-Farber/Harvard Cancer Center protocol 11-286). As
544 the trained foundational model is public, all the results can be reproduced using the accessible test datasets.

545

546 **CODE AVAILABILITY STATEMENT**

547 The complete pipeline used in this study can be accessed either from the [AIM webpage](#) or directly on [GitHub](#).
548 This includes the code for 1) Data download and pre-processing: Starting from downloading the data to
549 generating splits used in our study; 2) Replicating the training and inference of foundation and supervised
550 models across all tasks; and 3) Code for reproducing our comprehensive performance validation. In addition
551 to the code, we also provide trained model weights, extracted features, and outcome predictions for all the
552 models used in our study. Most importantly, we provide our foundation model accessible through a simple
553 pip package install and 2 lines of code to extract features for your data. We also provide a detailed
554 documentation website that can be accessed [here](#). The final model weights will also be made available through
555 the [Zenodo.org](#) platform as well as through [Mhub.ai](#) in a reproducible, containerized, off-the-shelf executable
556 format.

557

558

559 **COMPETING INTERESTS**

560 The authors declare no competing interests.

561

562 **REFERENCES**

- 563 1. Bommasani, R. *et al.* On the Opportunities and Risks of Foundation Models. *arXiv [cs.LG]* (2021).
- 564 2. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *arXiv [cs.CL]*
565 27730–27744 (2022).
- 566 3. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers
567 for Language Understanding. *arXiv [cs.CL]* (2018).
- 568 4. Radford, A. *et al.* Learning transferable visual models from natural language supervision. *arXiv [cs.CV]*
569 8748–8763 (18--24 Jul 2021).
- 570 5. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual
571 Representations. *arXiv [cs.LG]* (2020).
- 572 6. Oquab, M. *et al.* DINOv2: Learning robust visual features without supervision. *arXiv [cs.CV]* (2023).
- 573 7. Anja Thieme Microsoft Health Futures, United Kingdom *et al.* Foundation Models in Healthcare:
574 Opportunities, Risks & Strategies Forward. <https://doi.org/10.1145/3544549.3583177>
575 doi:10.1145/3544549.3583177.
- 576 8. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265
577 (2023).
- 578 9. Mahajan, A. *et al.* Deep learning-based predictive imaging biomarker model for EGFR mutation status in
579 non-small cell lung cancer from CT imaging. *J. Clin. Orthod.* **38**, 3106–3106 (2020).
- 580 10. Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics
581 study. *PLoS Med.* **15**, e1002711 (2018).
- 582 11. Braghetto, A., Marturano, F., Paiusco, M., Baiesi, M. & Bettinelli, A. Radiomics and deep learning
583 methods for the prediction of 2-year overall survival in LUNG1 dataset. *Sci. Rep.* **12**, 14132 (2022).
- 584 12. Balestrieri, R. *et al.* A Cookbook of Self-Supervised Learning. *arXiv [cs.LG]* (2023).
- 585 13. Huang, S.-C. *et al.* Self-supervised learning for medical image classification: a systematic review and

- 586 implementation guidelines. *NPJ Digit Med* **6**, 74 (2023).
- 587 14. Yan, K., Wang, X., Lu, L. & Summers, R. M. DeepLesion: automated mining of large-scale lesion
588 annotations and universal lesion detection with deep learning. *J Med Imaging (Bellingham)* **5**, 036501
589 (2018).
- 590 15. Zhao, B. *et al.* Evaluating variability in tumor measurements from same-day repeat CT scans of patients
591 with non-small cell lung cancer. *Radiology* **252**, 263–272 (2009).
- 592 16. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for Simplicity: The All Convolutional
593 Net. *arXiv [cs.LG]* (2014).
- 594 17. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding
595 noise. *arXiv [cs.LG]* (2017).
- 596 18. Hinshaw, D. C. & Shevde, L. A. The Tumor Microenvironment Innately Modulates Cancer Progression.
597 *Cancer Res.* **79**, 4557–4566 (2019).
- 598 19. Azizi, S. *et al.* Big Self-Supervised Models Advance Medical Image Classification. *arXiv [eess.IV]* (2021).
- 599 20. Krishnan, R., Rajpurkar, P. & Topol, E. J. Self-supervised learning in medicine and healthcare. *Nat*
600 *Biomed Eng* **6**, 1346–1352 (2022).
- 601 21. Ghesu, F. C. *et al.* Self-supervised Learning from 100 Million Medical Images. *arXiv [cs.CV]* (2022).
- 602 22. Haarbuerger, C. *et al.* Radiomics feature reproducibility under inter-rater variability in segmentations of
603 CT images. *Scientific Reports* vol. 10 Preprint at <https://doi.org/10.1038/s41598-020-69534-6> (2020).
- 604 23. Campello, V. M. *et al.* Minimising multi-centre radiomics variability through image normalisation: a pilot
605 study. *Sci. Rep.* **12**, 12532 (2022).
- 606 24. Kumar, D. *et al.* Discovery Radiomics for Pathologically-Proven Computed Tomography Lung Cancer
607 Prediction. *arXiv [cs.CV]* (2015).
- 608 25. Lao, J. *et al.* A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma
609 Multiforme. *Sci. Rep.* **7**, 10353 (2017).
- 610 26. Haarbuerger, C., Weitz, P., Rippel, O. & Merhof, D. Image-based Survival Analysis for Lung Cancer
611 Patients using CNNs. *arXiv [cs.CV]* (2018).
- 612 27. Cho, H.-H. *et al.* Radiomics-guided deep neural networks stratify lung adenocarcinoma prognosis from

- 613 CT scans. *Commun Biol* **4**, 1286 (2021).
- 614 28. Taleb, A. *et al.* 3d self-supervised methods for medical imaging. *Adv. Neural Inf. Process. Syst.* **33**,
615 18158–18172 (2020).
- 616 29. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest X-ray images via self-
617 supervised learning. *Nat Biomed Eng* **6**, 1399–1406 (2022).
- 618 30. Zhou, Z. *et al.* Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis. *Med. Image*
619 *Comput. Comput. Assist. Interv.* **11767**, 384–393 (2019).
- 620 31. Chaitanya, K., Erdil, E., Karani, N. & Konukoglu, E. Contrastive learning of global and local features for
621 medical image segmentation with limited annotations. *arXiv [cs.CV]* (2020).
- 622 32. Li, H. *et al.* Imbalance-Aware Self-supervised Learning for 3D Radiomic Representations. in *Medical*
623 *Image Computing and Computer Assisted Intervention – MICCAI 2021* 36–46 (Springer International
624 Publishing, 2021).
- 625 33. Li, Z. *et al.* A Novel Collaborative Self-Supervised Learning Method for Radiomic Data. *arXiv [eess.IV]*
626 (2023).
- 627 34. Zhao, Z. & Yang, G. Unsupervised Contrastive Learning of Radiomics and Deep Features for Label-
628 Efficient Tumor Classification. in *Medical Image Computing and Computer Assisted Intervention –*
629 *MICCAI 2021* 252–261 (Springer International Publishing, 2021).
- 630 35. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine Learning methods for
631 Quantitative Radiomic Biomarkers. *Sci. Rep.* **5**, 13087 (2015).
- 632 36. Adebayo, J., Gilmer, J. & Muelly, M. Sanity checks for saliency maps. *Adv. Neural Inf. Process. Syst.*
633 (2018).
- 634 37. Arun, N. *et al.* Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical
635 Imaging. *Radiol Artif Intell* **3**, e200267 (2021).
- 636 38. Setio, A. A. A. *et al.* Validation, comparison, and combination of algorithms for automatic detection of
637 pulmonary nodules in computed tomography images: The LUNA16 challenge. *Med. Image Anal.* **42**, 1–
638 13 (2017).
- 639 39. Kirby, J. NSCLC-Radiomics. <https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics>.

- 640 40. Napel, S. NSCLC radiogenomics: Initial Stanford study of 26 cases. *The Cancer Imaging Archive*.
- 641 41. Wang, F. & Liu, H. Understanding the behaviour of contrastive loss. *arXiv [cs.LG]* 2495–2504 (2020).
- 642 42. Uemura, T., Näppi, J. J., Hironaka, T., Kim, H. & Yoshida, H. Comparative performance of 3D-DenseNet,
643 3D-ResNet, and 3D-VGG models in polyp detection for CT colonography. in *Medical Imaging 2020:
644 Computer-Aided Diagnosis* vol. 11314 736–741 (SPIE, 2020).
- 645 43. Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Adv. Neural Inf. Process.
646 Syst.* **29**, (2016).
- 647 44. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]* 2825–2830 (2012).
- 648 45. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter
649 Optimization Framework. in *Proceedings of the 25th ACM SIGKDD International Conference on
650 Knowledge Discovery & Data Mining* 2623–2631 (Association for Computing Machinery, 2019).
- 651 46. Gmail, L. & Hinton, G. Visualizing Data using t-SNE.
652 <https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbclid=...> (2008).
- 653 47. Jorge Cardoso, M. *et al.* MONAI: An open-source framework for deep learning in healthcare. *arXiv
654 [cs.LG]* (2022).
- 655 48. Gamer, M. irr : Various Coefficients of Interrater Reliability and Agreement. [http://cran.r-
656 project.org/web/packages/irr/irr.pdf](http://cran.r-project.org/web/packages/irr/irr.pdf) (2010).

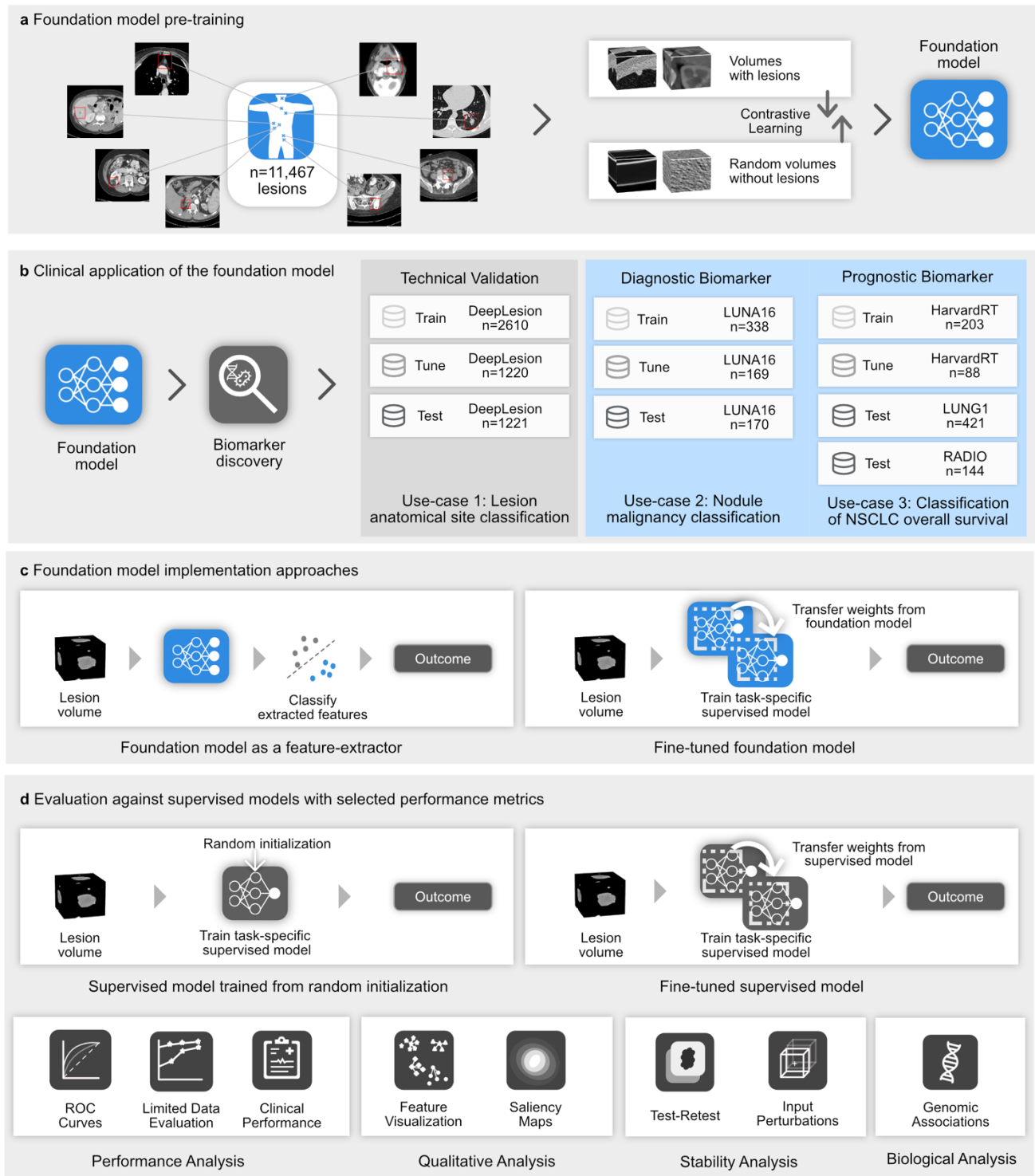
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671

672

673

674

FIGURES



675

676

677

678

679

680

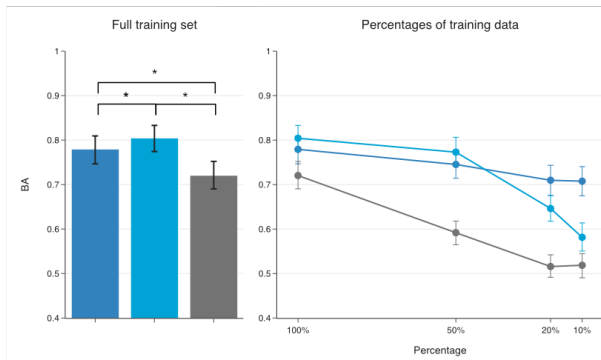
681

682

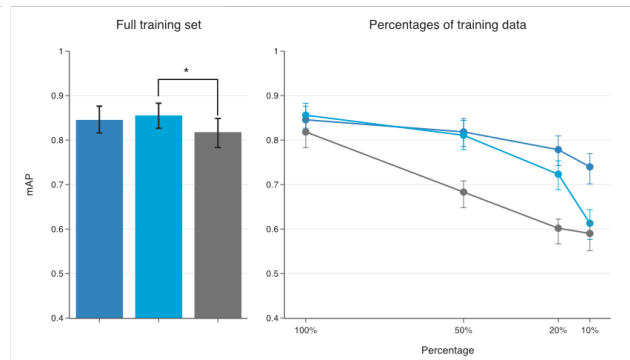
Figure 1 | General overview of the study. a. Foundation model pre-training. A foundation model, specifically a deep convolutional encoder model, was pre-trained by contrasting volumes with and without lesions. **b. Clinical application of the foundation model.** The foundation model was used to extract biomarkers and then evaluated for three classification tasks on diverse datasets. **c. Foundation model implementation approaches** The foundation model was adapted to specific use cases by extracting features or through fine-tuning (left). **d. Evaluation against supervised models with selected performance metrics.** We compared the performance of the foundation models against conventional supervised implementations, trained from random initialization (left) and fine-tuned from a different task (right). The comparison was made through several criteria for the different use

683 cases, including quantitative performance, stability, and biological analysis. Biological, clinical, and stability analyses are limited to use case 2 due to
 684 the availability of associated data.

a Balanced accuracy of lesion anatomical site classification for the full training set and percentages of training data

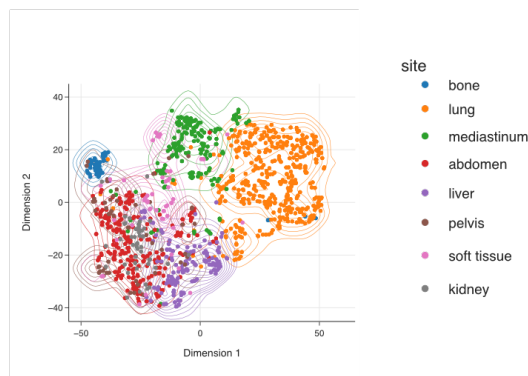


b Mean average precision of lesion anatomical site classification for the full training set and percentages of training data

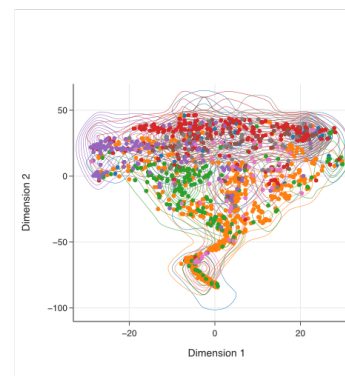


Models ■ Foundation model (Features) ■ Foundation model (Finetuned) ■ Supervised (Random init)

c Visualization of features extracted from the foundation model after t-SNE dimensionality reduction



d Visualization of features extracted from the supervised model after t-SNE dimensionality reduction



685

686

687 **Figure 2 | Performance of foundation model for lesion anatomical site classification.** We compared foundation model adaptation approaches
 688 against a supervised model using balanced accuracy (a) and mean average precision (b). We show performance on these metrics computed across the
 689 eight anatomical sites for the full training set and when the training data percentage is decreased to 50%, 20%, and 10%. Error bars in (a) and (b)
 690 show 95% confidence intervals of the estimates. Visual representation of the features generated from the independent test-set for identifying lesion
 691 anatomical sites, using c the foundation model as a feature extractor, and d the supervised model. For (c) and (d), the x-axis corresponds to dimension
 692 1, and the y-axis to dimension 2 of the t-SNE dimensionality reduction. The density contours belonging to each class are overlaid for (c) and (d) to
 693 highlight separability between classes in the feature space.

694

695

696

697

698

699

700

701

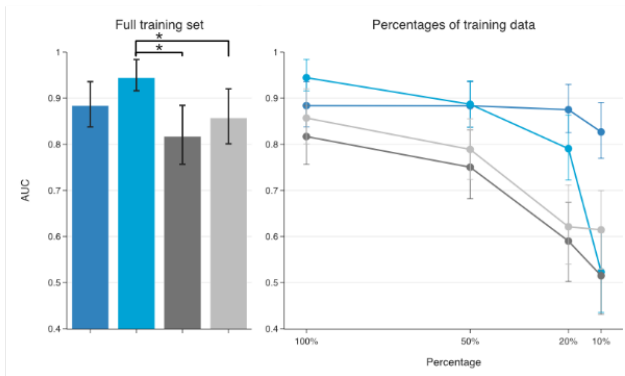
702

703

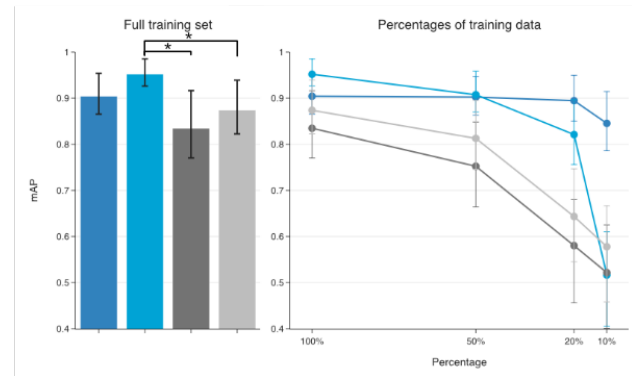
704

705

a Area under the receiver operating curve of nodule malignancy classification for full training set and percentages of training data

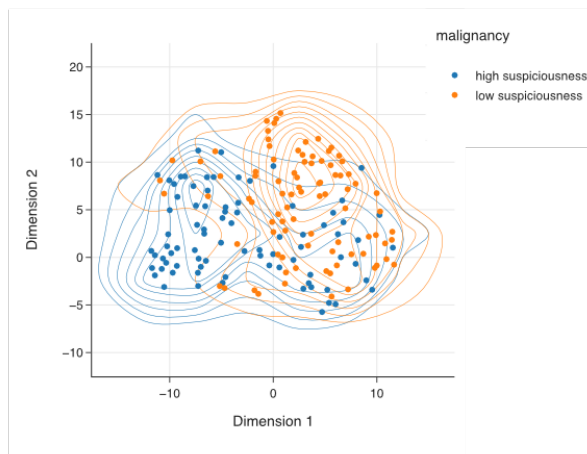


b Average precision of nodule malignancy classification for full training set and percentages of training data

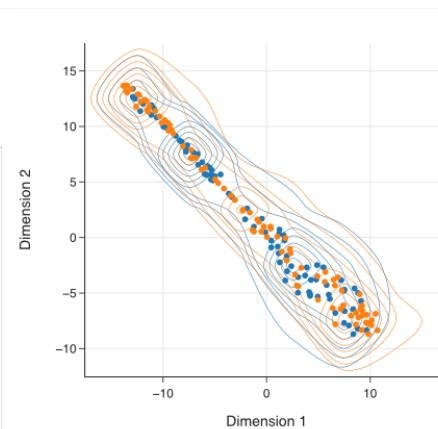


Models ■ Foundation model (Features) ■ Foundation model (Finetuned) ■ Supervised (Random init) ■ Supervised (Finetuned)

c Visualization of features extracted from the foundation model after t-SNE dimensionality reduction



d Visualization of features extracted from the fine-tuned supervised model after t-SNE dimensionality reduction



706

707

708

709

710

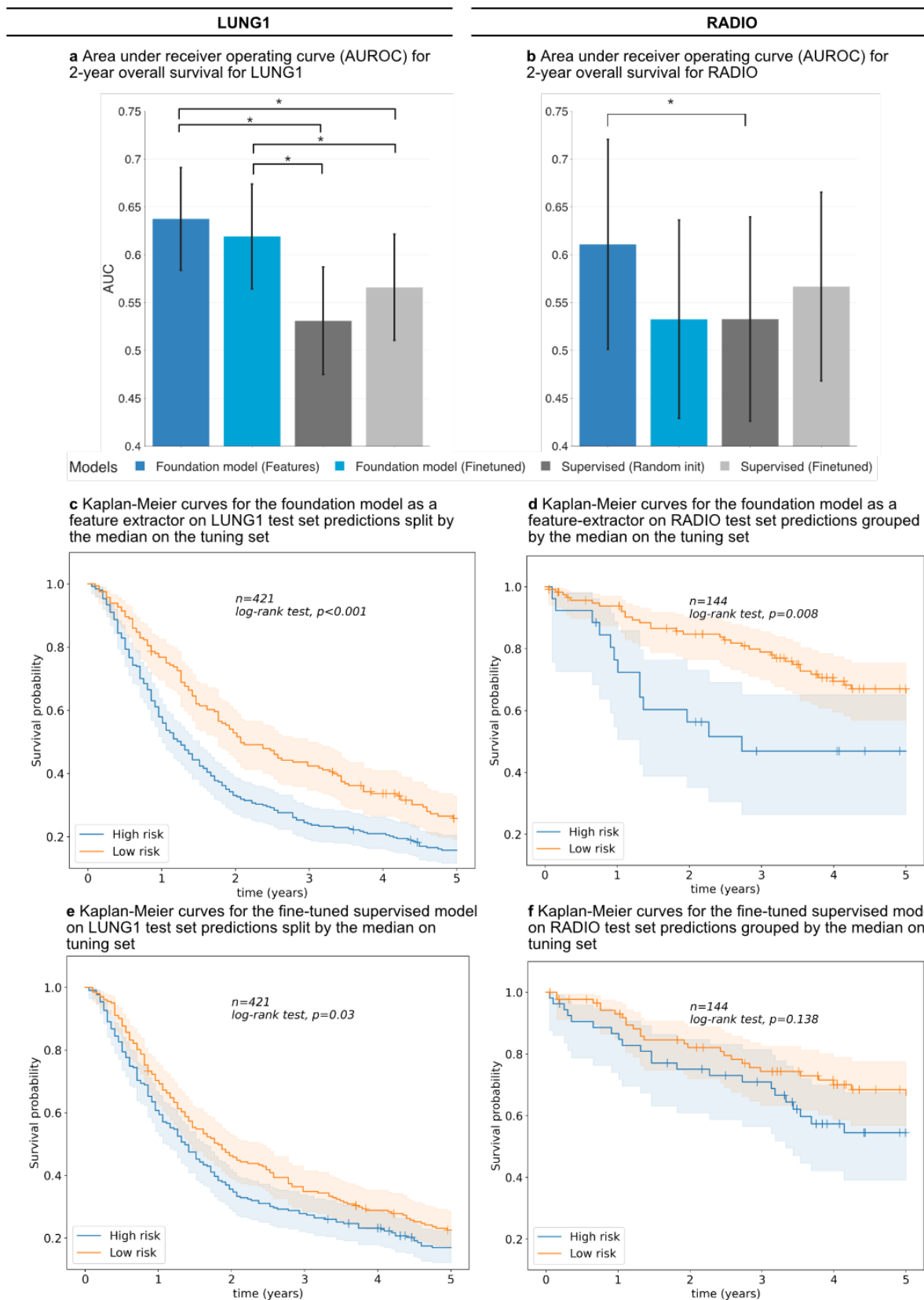
711

712

713

714

Figure 3 | Performance comparison of the foundation model against supervised for nodule malignancy prediction. We compared the foundation model adaptation approaches against baseline supervised models using the full training dataset and on decreasing the training data percentages to 50%, 20% and 10%. **a** Area under receiver operating curves (AUC-ROC) **b** Average precision (AP). Error bars in **(a)** and **(b)** show 95% confidence intervals of the estimates. Visual representation of the features generated from the independent test-set for the task of nodule malignancy prediction using, **c** the fine-tuned supervised model and **d** using the foundation model as a feature extractor. For **(c)** and **(d)**, the x-axis corresponds to dimension 1, and the y-axis to dimension 2 of the t-SNE dimensionality reduction. The density contours belonging to each class are overlaid for **(c)** and **(d)** to highlight separability between classes in the feature space.



715

716

717

718

719

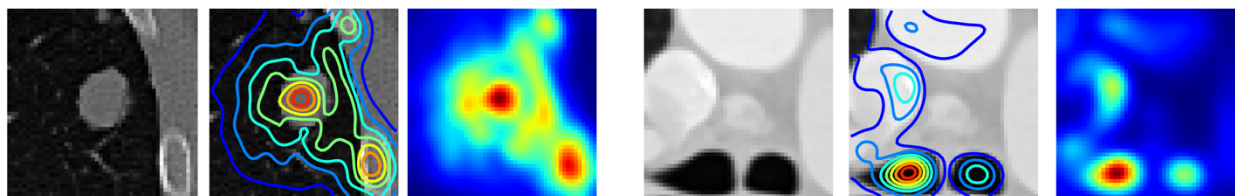
720

721

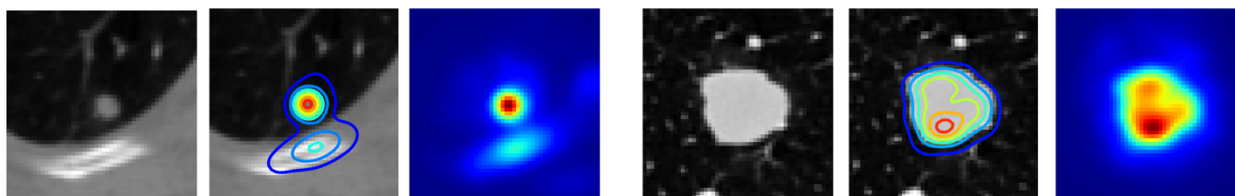
722

Figure 4 | Performance of the foundation model against supervised for prognostication of NSCLC tumors. We compared the foundation model against the baseline supervised models using the area under the receiver operating curve (AUC) for 2-year overall survival for **a** LUNG1 **b** RADIO. Kaplan-Meier (KM) curves for predictions generated from the foundation model as a feature-extractor for LUNG1 (**c**) and RADIO (**d**) as well as the fine-tuned supervised method for LUNG1 (**e**) and RADIO (**f**). To ensure a fair comparison, we calculated the threshold for the split between the KM groups on the tuning set for each network. Kaplan-Meier curves for the other approaches, fine-tuning the foundation model and training a supervised model from random initialization can be found in Fig. S1 in the supplementary.

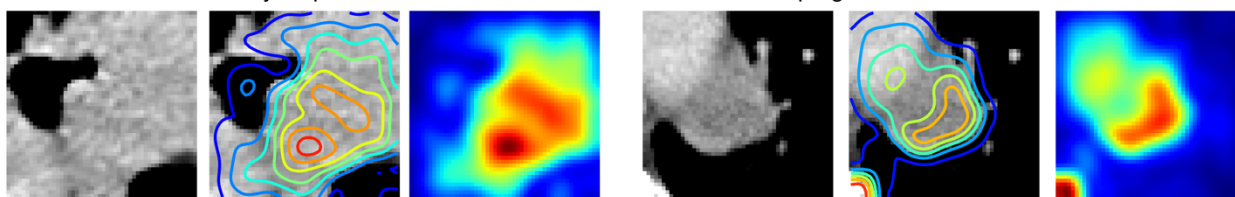
a Gradient-based saliency map of the foundation model fine-tuned for lesion anatomical site classification



b Gradient-based saliency map of the foundation model fine-tuned for malignancy prediction



c Gradient-based saliency map of the foundation model fine-tuned for cancer prognostication



723

724

725

726

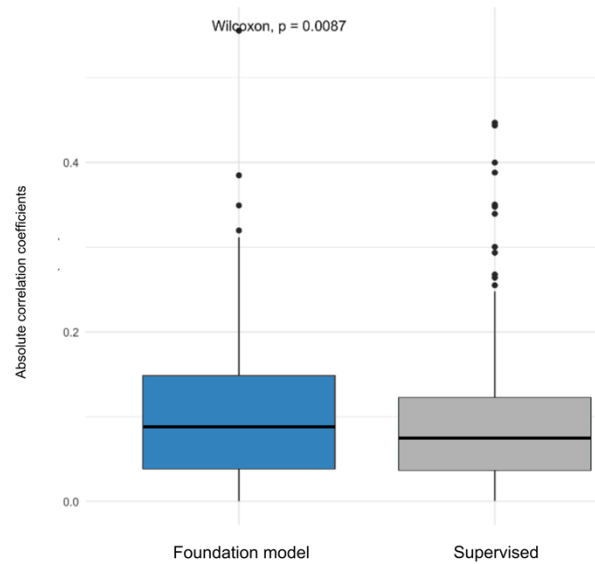
727

728

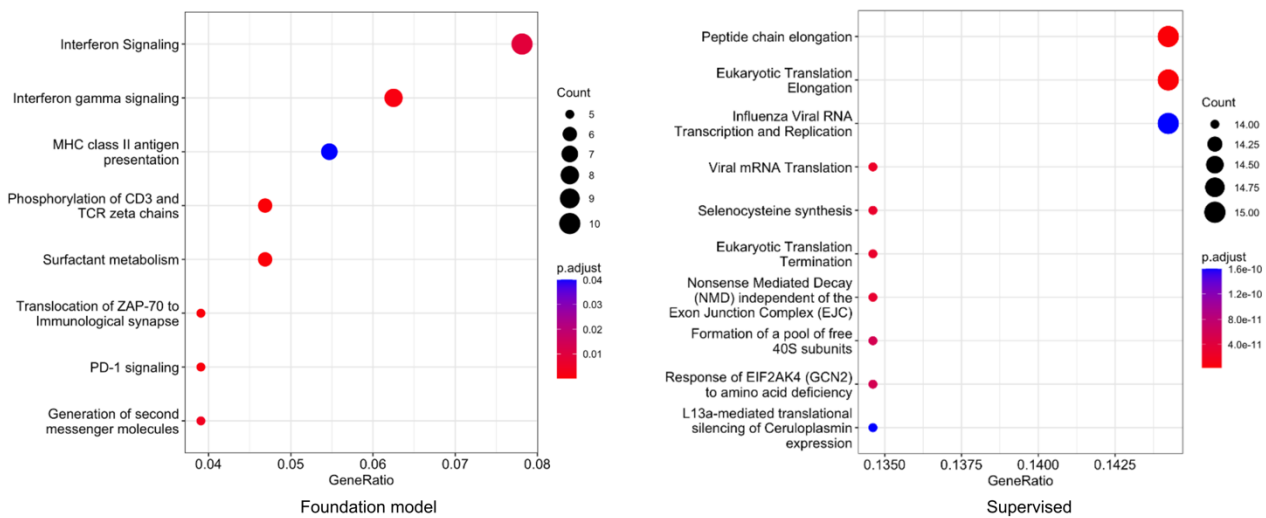
729

Figure 5 | Saliency maps for fine-tuned foundation models. We generated gradient-based saliency maps for each of the fine-tuned foundation models from use cases I (a), II (b), and III (c) using smooth guided backpropagation and visualized salient regions on two samples from corresponding test datasets. The first and fourth columns show the central axial slice (50mm x 50mm) of the volume provided as input to the self-supervised network. The second and fifth columns show isolines for saliency contours. Finally, the third and sixth columns show saliency maps highlighting areas of the input volume that contribute the most to a change in the output prediction.

a Absolute correlation coefficients between gene expression profiles and predictions of the feature-extractor foundation and fine-tuned supervised models



b Gene-set enrichment analysis to describe genes associated with each of the feature-extractor foundation and fine-tuned supervised model predictions. Genes with a correlation co-efficient > 0.1 were selected for the analysis



730
731

732 **Figure 6 | Underlying biological basis of the foundation model.** We compared the foundation and supervised model predictions with gene expression
733 profiles. **a** Box plot of absolute correlation coefficients (y-axis) of selected genes against model predictions (x-axis). **b** Gene-set enrichment analysis of
734 genes with correlation coefficient > 0.1 revealed for the foundation (left) and supervised model predictions (right). Genetic pathways are shown on the
735 y-axis, and the gene ratio is shown on the x-axis. Gene count and adjusted p-values are also shown in the legend.

736
737
738
739
740
741
742
743

744 **EXTENDED DATA**

745

Foundation Model Implementation	Data percentage	Increase in BA over supervised (95% CI, p-value)	Increase in mAP over supervised (95% CI, p-value)
Feature-extractor	50% (n=2526)	0.153 (0.123-0.186, p<0.005)	<i>0.135</i> (0.104-0.168, p<0.005)
Fine-tuned		<i>0.181</i> (0.147-0.214, p<0.005)	0.127 (0.097-0.162, p<0.005)
Feature-extractor	20% (n=1010)	<i>0.194</i> (0.159-0.228, p<0.005)	<i>0.177</i> (0.142-0.216, p<0.005)
Fine-tuned		0.130 (0.102-0.159, p<0.005)	0.121 (0.089-0.159, p<0.005)
Feature-extractor	10% (n=505)	<i>0.189</i> (0.148-0.228, p<0.005)	<i>0.149</i> (0.112-0.189, p<0.005)
Fine-tuned		0.063 (0.028-0.098, p<0.005)	0.02 (-0.011- 0.061, p=0.28)

746

747 **Extended Data Table 1 | Detailed comparison of the foundation model implementations against supervised methods in limited data settings for**
 748 **lesion anatomical site classification** Comparison of the foundation model as a feature-extractor and fine-tuned against the randomly initialised
 749 supervised model at 50%, 20% and 10% training data. For each data percentage, the largest increase in performance between the two is shown italicised.
 750 Not significant results are shown in red

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

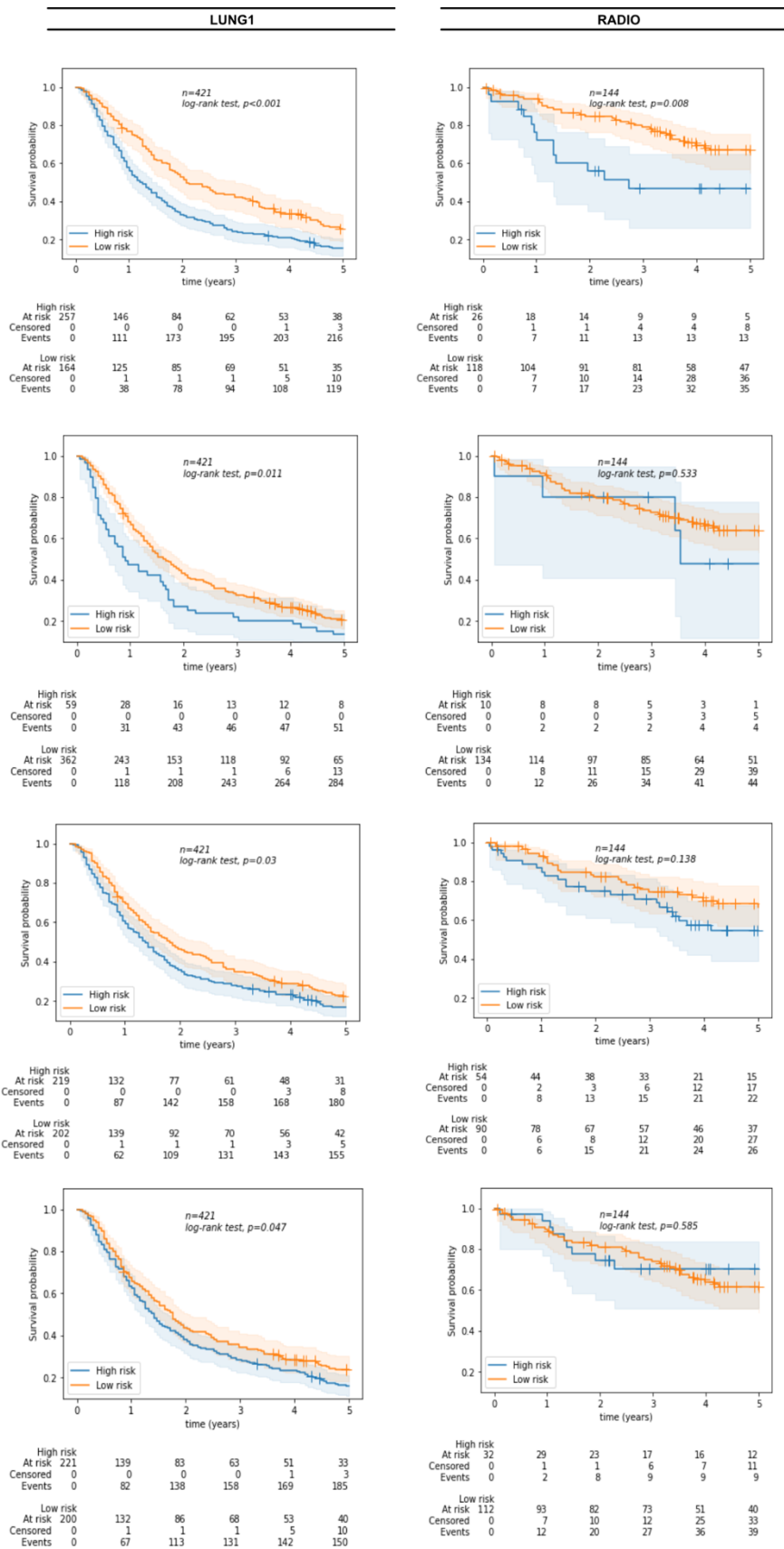
770

771
772
773
774

Foundation Model Implementation	Data percentage	Increase in AUC over supervised random initialization (95% CI, p-value)	Increase in mAP over supervised random initialization (95% CI, p-value)	Increase in AUC over supervised fine-tuned (95% CI, p-value)	Increase in mAP over supervised fine-tuned (95% CI, p-value)
Feature-extractor	50% (n=254)	0.133 (0.064 - 0.207, p<0.005)	0.15 (0.068 - 0.222, p<0.005)	0.07 (0.021 - 0.167, p<0.05)	0.089 (0.024 - 0.153, p=0.063)
Fine-tuned		0.136 (0.070-0.199, p<0.005)	0.155 (0.083-0.223, p<0.005)	0.097 (0.035-0.155366, p<0.05)	0.095 (0.035-0.148, p<0.005)
Feature-extractor	20% (n=101)	0.285 (0.193-0.370, p<0.05)	0.314 (0.227-0.420, p<0.005)	0.254 (0.173-0.330, p<0.05)	0.251 (0.164-0.334, p<0.005)
Fine-tuned		0.20 (0.092-0.308, p<0.005)	0.24 (0.138-0.35, p<0.005)	0.169 (0.093-0.245, p<0.005)	0.177 (0.089-0.260, p<0.005)
Feature-extractor	10% (n=51)	0.312 (0.211-0.408, p<0.005)	0.323 (0.238-0.423, p<0.005)	0.212 (0.128-0.285, p<0.005)	0.268 (0.179-0.376, p<0.005)
Fine-tuned		0.008 (-0.089 -0.101, p=0.919)	-0.005 (-0.095-0.08, p=0.869)	-0.091 (-0.015 - 0.171481, p<0.05)	-0.061 (-0.144 - 0.023, p=0.322)

775
776
777
778
779
780

Extended Data Table 2 | Detailed comparison of the foundation model implementations against supervised methods in limited data settings for nodule malignancy classification Comparison of the foundation model as a feature-extractor and fine-tuned against randomly initialised and fine-tuned supervised models at 50%, 20% and 10% of the training data. For each data percentage, the largest increase in performance between the two is shown italicised. Not significant results are shown in red



781

782

783 **Extended Data Figure 1 | Kaplan Meier curves for all models investigated** Kaplan Meier curves for the LUNG1 and RADIO datasets for the foundation

784 model as a feature-extractor (first row), fine-tuned foundation model (second row), fine-tuned supervised model (third row) and randomly initialised

785 supervised model (last row)

786

787

	LUNG1			RADIO		
	beta	HR (95% CI for HR)	p.value	beta	HR (95% CI for HR)	p.value
Foundation model as feature extractor	-0.44	0.65 (0.52-0.81)	<0.005	-0.84	0.43 (0.23-0.82)	0.01
Foundation model fine-tuned	-0.39	0.68 (0.5-0.92)	0.01	-0.32	0.72 (0.26-2.01)	0.53
Supervised (fine-tuned)	-0.24	0.79 (0.64-0.98)	0.03	-0.43	0.65 (0.37-1.15)	0.14
Supervised (random initialization)	-0.22	0.80 (0.65-1.00)	0.05	0.20	1.22 (0.59-2.53)	0.59

788

789

790 **Extended Data Table 3 | Univariate cox regression** Results of univariate cox models showing the relationship between implementations of the
 791 foundation model and the supervised methods and survival on LUNG1 and RADIO datasets. The median split on the training dataset (HarvardRT) is used,
 792 also shown in Fig S4 in the Kaplan-Meier curves.

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

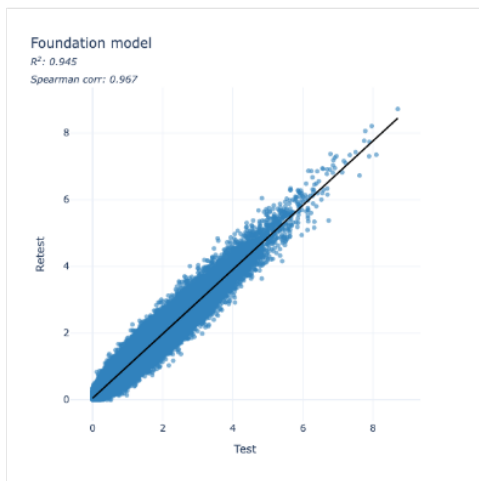
814

815

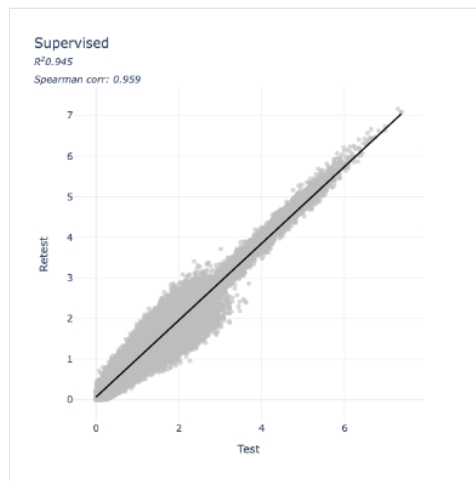
816

817

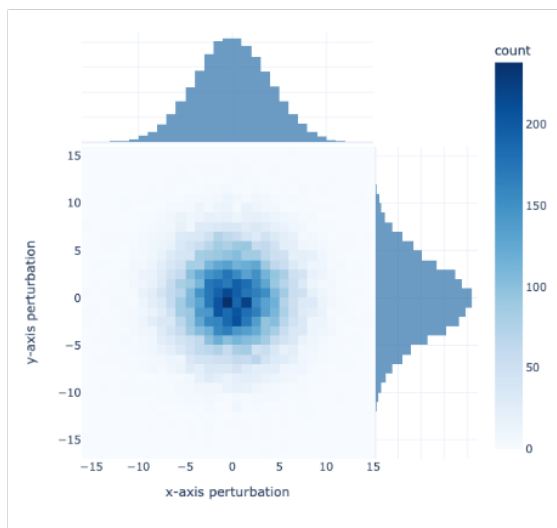
a RIDER test features against retest features for the foundation model as a feature-extractor



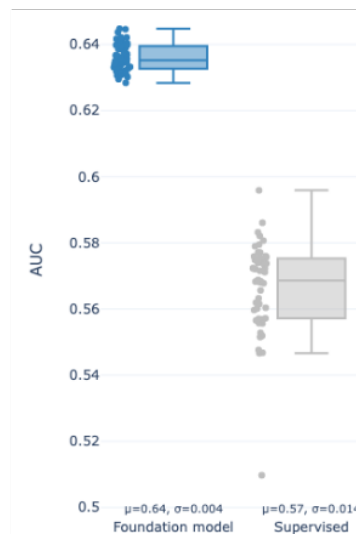
b RIDER test features against retest features for the fine-tuned supervised model



c Sampling distribution for input perturbations



d Input stability of AUC for 2-year survival for the feature-extractor foundation and fine-tuned supervised model



818

819

820

821

822

823

824

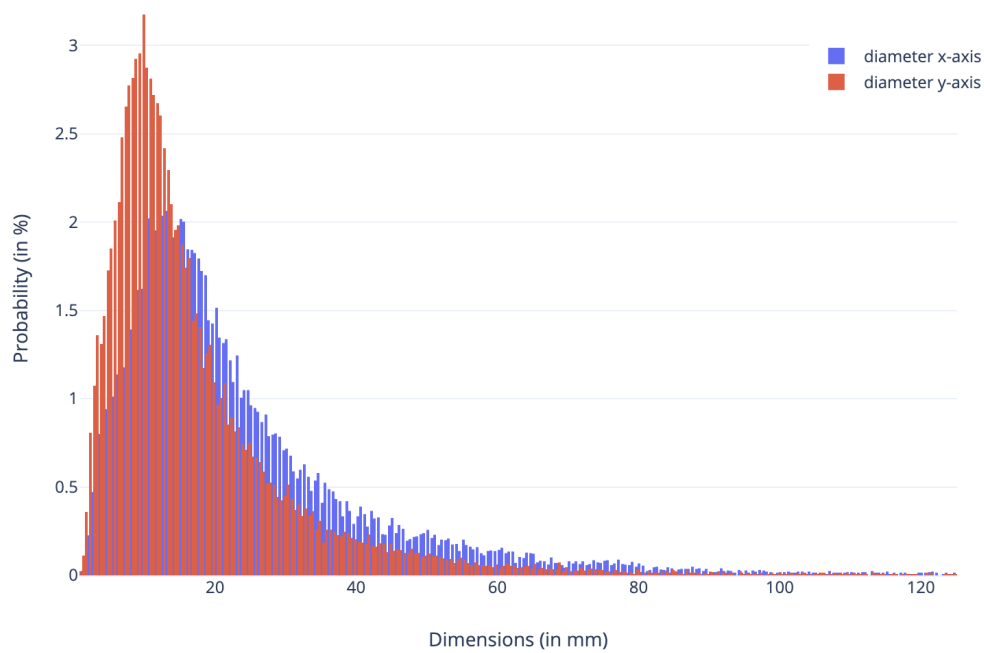
825

826

827

Extended Data Figure 2 | Stability of self-supervised learning networks. We analyzed the test-retest robustness on the RIDER dataset by comparing the correlation between features generated by **a.** the foundation model as a feature extractor and **b.** the fine-tuned supervised model. In **c.**, the inter-reader variability is simulated by adding perturbations from a sampling distribution. We perturb across x, y and z-axes although the distribution is shown only for x and y perturbations for simplicity. **d** Prognostic stability of the feature extractor foundation model against the fine-tuned supervised model when the input seed point is perturbed, estimated through AUC for 2-year survival.

Lesion Dimensions



828
829

830 **Extended Data Figure 3 | Diameter distribution of DeepLesion** Distribution of diameters in the x and y axes for the DeepLesion training dataset based
831 on RECIST bookmarks identified on key slices. Input dimensions of 50x50x50 mm³ were chosen as they covered 93% and 97% of the distribution in the x
832 and y axes respectively.

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

	Pre-training	Use-case 1: Lesion Anatomical Site Classification			Use-case 2: Nodule Malignancy Classification			Use-case 3: Classification of survival for NSCLC tumors				Stability
Cohorts	DeepLesion	DeepLesion			LUNA16			HarvardRT	LUNG1	RADIO	RIDER	
Institution	NIH Clinical Center	NIH Clinical Center			Multi-center			Dana-Farber Cancer Center	MAASTRO Clinic	Stanford & Palo Alto VA	MSKCC	
Usage	Pre-train	Train	Tune	Test	Train	Tune	Test	Train	Tune	Test	Test	Test
Scans	11,467	2610	1220	1221	338	169	170	203	88	421	144	52
Patients	2,312	553	379	390	266	149	150	203	88	421	144	26

854

		Use-case 1: Lesion Anatomical Site Classification		Use-case 2: Nodule Malignancy Classification		Use-case 3: Classification of survival for NSCLC tumors					
						HarvardRT		LUNG1		RADIO	
Outcome Distribution	bone	4.1%	benign	51.7%	alive (2-year)	54.2%	alive (2-year)	59.8%	alive (2-year)	64.5%	
	abdomen	16.3%									
	mediastinum	14.3%									
	liver	9.7%									
	lung	41.1%	malignant	48.3%	dead (2-year)	45.7%	dead (2-year)	40.1%	dead (2-year)	35.4%	
	kidney	3.6%									
	soft tissue	4.6%									
	pelvis	6.0%									
Sex	M	58.5%	na		52.2%		68.8%		75%		
	F	41.5%	na		47.7%		31.1%		25%		
Age (median)	58.0		na		69.6		68.69		69.0		

855

856

857 **Extended Data Table 4 | Dataset breakdown** Table showing the 6 different cohorts used in this study along with eligible scans and patients used. A
 858 secondary table shows the outcome, sex, and age distribution of each of the cohorts.

859

860

861

862

863