Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 2334–2344
https://doi.org/10.54941/ahfe1008459

When the Model Card is the Hazard: A Silent Preprocessing Failure in a CT Foundation Model and a Label-Free Check for Trustworthy Deployment

Woo Hyun Shim[1], Chae Ri Park[1], Sang Cheol Kim[2], and Sukyoung An[2]

[1]Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul 05505, Republic of Korea
[2]Division of Healthcare and Artificial Intelligence, Department of Precision Medicine, National Institute of Health, Korea Disease Control and Prevention Agency, Cheongju 28159, Republic of Korea

ABSTRACT

Clinical teams increasingly build decision support on frozen foundation models (FMs), preparing inputs as the public model card shows. We report a case in which the documentation itself was the hazard. A linear probe on CT-FM embeddings was trained to predict COVID-19 outcome in 1,205 patients and applied unchanged to 309 patients at another hospital. Following the model card, external mortality AUC (area under the receiver operating characteristic curve) was 0.64; restoring the resampling used in pretraining, documented only in the training code, gave 0.71 (0.73 at 2 mm; Holm $P = .10$). In 319 examinations reconstructed both thin and thick, the model-card pipeline raised predicted risk by 1.76 standard deviations (SD) and scrambled patient ranking (Spearman 0.34): a thicker slice looked like a sicker patient. The corrected pipeline was stable (−0.02 SD; 0.96). The failure was reproduced by removing resampling from two other FMs on 1,110 CTs from a third cohort, and model-side repairs did not fully fix it. A Korean cohort (255 patients) analysed under a prospectively specified plan added kernel-driven instability and, post hoc, inflated predicted risk. We frame the omission as a latent condition that is invisible to clinicians and unequal across sites, and propose a three-step pre-deployment check that clinical teams can run: reproduce the pretraining preprocessing, test stability on paired reconstructions without labels, and recalibrate on about ten local cases.

Keywords: Trustworthy AI, Model documentation, Foundation models, Medical imaging, Deployment safety, Human factors in healthcare, Acquisition bias

INTRODUCTION

Imaging artificial intelligence (AI) for COVID-19 largely failed on external validation, often because models learned shortcuts linked to how or where images were acquired rather than to disease, as shown earlier for pneumonia (Zech et al., 2018) and for COVID-19 (DeGrave, Janizek and Lee, 2021; Roberts et al., 2021). A newer generation of three-dimensional CT foundation models (FMs), pretrained on large unlabelled image collections, now offers

Received October 29, 2026; Revised November 14, 2026; Accepted November 20, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

frozen embeddings: fixed numerical summaries of a scan that a hospital team can feed into a small predictor trained on a few hundred labelled patients. These embeddings are increasingly used as drop-in features. The steps that prepare a scan before it enters the model (orientation, intensity window, cropping and resampling to a fixed voxel size) are usually copied from the model card, the public documentation released with a model (Mitchell et al., 2019).

From a human-factors perspective, the model card is an interface between the developers of a model and the clinical team that deploys it. If that interface is incomplete, the resulting error does not look like anyone's mistake. A user who follows the documentation exactly is doing the right thing; the flaw was built into the system upstream and lies dormant until it meets local conditions, which is what Reason (2000) calls a latent condition rather than an active error. CT measurements are known to depend on slice thickness (the spacing between consecutive cross-sectional images) and on the reconstruction kernel (the soft or sharp filter used to compute images from raw scanner data) (Lu et al., 2016; Park et al., 2025). Those studies measured feature reproducibility. Whether a mismatch between documented and actual preprocessing can masquerade as a model deficiency has not been tested by intervention.

We asked three questions. (1) When a frozen FM's external prognostic performance collapses, is the cause the model, the cohort or the input preparation? (2) Can the cause be established by changing one thing at a time while holding the weights, the predictor and the patients fixed? (3) What can a clinical team check before deployment, without outcome labels? We treated an observed transfer failure as a hypothesis and tested it with controlled preprocessing interventions, real paired reconstructions, a third public cohort and a Korean cohort analysed under a prospectively specified plan.

METHODS

Data and Governance

The training cohort was the STOIC2021 public set: 1,205 patients positive by reverse-transcription polymerase chain reaction (RT-PCR) with the challenge outcome of intubation or death within 1 month (prevalence 25%) (Revel et al., 2021; Boulogne et al., 2024). The external test cohort was COVID-19-NY-SBU (Saltz et al., 2021): 309 patients with CT and in-hospital outcome (54 deaths), one series per patient, non-contrast preferred. In this cohort, 333 acquisitions from 289 patients had both a thin ($\leq$ 1.5 mm, mostly 1.25 mm) and a thick ($\geq$ 4.5 mm, mostly 5 mm) reconstruction of the same scan, forming real same-acquisition pairs; 155 pairs shared the reconstruction kernel. MosMedData (Morozov et al., 2020) provided 1,110 CTs at 8-mm spacing with visual extent grades CT-0 to CT-4 (not prognosis). Median slice spacing was 0.8 mm in STOIC2021, 3 mm in NY-SBU (series up to 7 mm), 8 mm in MosMedData and 3 mm in the Korean cohort.

The Korean cohort comprised 255 patients with 436 visits and 3-mm axial reconstructions (427 of 431 main series); 241 patients had a first-visit

CT and a US National Institutes of Health (NIH) COVID-19 severity grade recorded in the clinical data (Severe or Critical, 42; 240 with lung embeddings). It contained 378 same-acquisition pairs reconstructed with a soft and a sharp kernel and 179 consecutive-visit pairs. Series selection rules, hypotheses and decision criteria were fixed in a time-stamped analysis plan before any foundation-model embeddings were computed for this cohort. Public datasets were used under their respective licences. The Korean cohort was de-identified before analysis and analysed within a Korea National Institute of Health data project on a national AI computing facility, under the Institutional Review Board approval of the Korea Disease Control and Prevention Agency for that project, in which the authors are registered researchers.

Models, Probe and Interventions

Three publicly released 3D CT FMs were used with frozen weights. For CT-FM (Pai et al., 2025), the model card's feature-extraction example applies orientation, an intensity window of −1024 to 2048 Hounsfield units (HU) and foreground cropping, with no resampling; the pretraining configuration in the public repository resamples to $3 \times 1 \times 1$ mm (z, y, x). We call the former the model-card pipeline and the latter the pretraining preprocessing. COLIPRI (Wald et al., 2025) inputs were resampled to 2 mm isotropic and cropped to 192^3 voxels centred on the lung mask (the released processor crops at the image centre); Merlin (Blankemeier et al., 2026), pretrained on abdominal CT, resamples to $1.5 \times 1.5 \times 3$ mm (x, y, z). Embeddings were pooled over the whole volume (global) and within an automatic lung mask (lung); masks agreed across paired reconstructions (thick/thin volume ratio, median 1.003; $r = 0.95$).

The probe standardised the embedding, reduced it to 32 principal components, appended age and sex, and fitted L2-penalised logistic regression ($C = 0.05$). It was trained once on STOIC2021 and never refitted on target data. A reference model used age, sex and seven quantitative CT measures (consolidation and ground-glass fractions, lung attenuation statistics, lung volume). All interventions changed only input preprocessing. For CT-FM we compared the model-card pipeline (native spacing), the pretraining preprocessing, isotropic 1-mm and 2-mm resampling, and component changes (z-only, in-plane-only, intensity window only). For COLIPRI and Merlin we removed z-resampling (reverse interventions). Synthetic thickening re-rendered thin NY-SBU series as 3-mm and 5-mm slabs with fixed field of view, and a post-hoc in-plane blur ($\sigma = 1$ mm) applied to training and target inputs tested whether kernel sensitivity could be removed at the input.

Outcome Measures and Statistics

On paired inputs we measured three things. Risk shift is the mean difference in predicted logit between the two versions of the same scan, in units of the standard deviation (SD) of the first version: does everyone's score move? Ranking agreement is the Spearman correlation of the two sets of logits: are the same patients still ranked as highest risk? The paired difference in

area under the receiver operating characteristic curve (AUC) asks whether discrimination of mortality changes. Three primary hypotheses were declared before the final analysis: H1, COLIPRI (lung) versus quantitative CT; H2 and H3, CT-FM with 2-mm resampling versus native spacing for lung and global pooling. They were tested with a double bootstrap that resampled both training and test patients and refitted the probe in each of 2,000 replicates, with Holm correction; secondary double-bootstrap comparisons used 500 replicates and all others a test-set paired bootstrap (1,000), and these are secondary or exploratory. Calibration was summarised by slope and Brier score (Van Calster et al., 2019), and local recalibration refitted the intercept, or intercept and slope, on n labelled target patients. In the Korean cohort, decision rules were fixed in advance for H-K1 (CT-FM lung 2 mm versus native for Severe or Critical versus other grades; supported if the lower confidence bound exceeded 0, refuted if the upper bound was below 0.03) and H-K2 (lower kernel-pair ranking for native than resampled CT-FM). The pre-specified site-cluster bootstrap for H-K1 was replaced post hoc by a patient-level double bootstrap because all but one of the evaluable patients came from a single site.

RESULTS

Following the Model Card Degraded External Prediction

Applied unchanged to NY-SBU (Table 1), the STOIC2021-trained probe reached a mortality AUC of 0.77 with COLIPRI lung embeddings, 0.76 with Merlin and 0.70 with quantitative CT, but 0.64 with CT-FM processed as in its model card. Restoring the pretraining preprocessing raised CT-FM to 0.71, and isotropic 2-mm resampling to 0.73 (H2, +0.091; 95% CI 0.006, 0.197; uncorrected P = .034; Holm P = .10). AUC rose in steps with resampling (native 0.64, 1 mm 0.71, 2 mm 0.73; Figure 1a). The 2-mm gain was positive in all 20 probe settings (8–128 principal components × C 0.01–1) and under two alternative per-patient series rules. With training-set variability included, COLIPRI's advantage over quantitative CT was not established (H1, Holm P = .31).

Table 1: External prediction of in-hospital death on NY-SBU (309 patients, 54 deaths) by a probe trained on STOIC2021 and applied unchanged. Differences are point estimates from the double bootstrap (2,000 replicates; probe refitted on the 1,204 STOIC2021 patients with all embeddings), so they need not equal differences of the AUCs above.

Input (Pooling, Preprocessing)	AUC [95% CI]	n
Quantitative CT (reference)	0.705 [0.625, 0.777]	309
COLIPRI (lung, 2 mm isotropic)	0.773 [0.703, 0.839]	309
Merlin (global, 1.5 × 1.5 × 3 mm)	0.762 [0.686, 0.830]	309
CT-FM (lung), model-card pipeline	0.636 [0.558, 0.704]	309

(Continued)

Table 1: Continued.

Input (Pooling, Preprocessing)	AUC [95% CI]	n
CT-FM (lung), pretraining preprocessing	0.708 [0.624, 0.780]	308
CT-FM (lung), 2 mm isotropic	0.734 [0.658, 0.803]	309
Pre-declared hypothesis	Difference [95% CI]	Holm P
H1: COLIPRI lung – quantitative CT	+0.070 [−0.022, 0.132]	.31
H2: CT-FM lung, 2 mm – model card	+0.091 [0.006, 0.197]	.10
H3: CT-FM global, 2 mm – model card	+0.051 [−0.059, 0.135]	.44

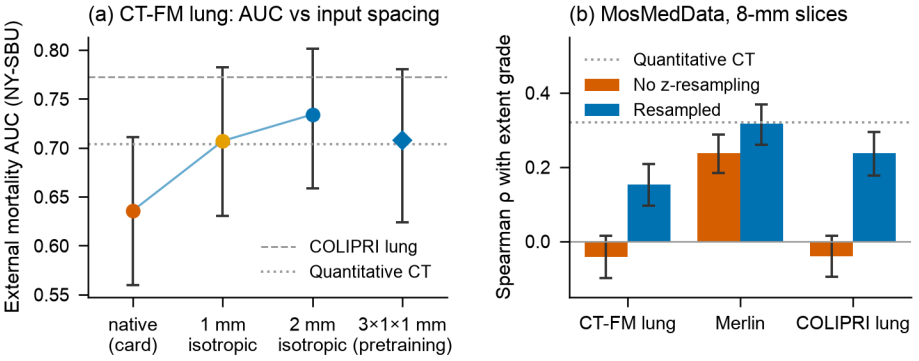


Figure 1: External mortality AUC of the CT-FM lung probe on NY-SBU by input spacing (a), with COLIPRI and quantitative CT as references, and replication on MosMedData 8-mm CTs (b): Spearman correlation with visual extent grade with and without z-resampling in three models. Bars and whiskers show estimates and 95% bootstrap confidence intervals.

A Thicker Slice Looked Like a Sicker Patient

The most direct evidence came from scans reconstructed both thin and thick (Table 2, Figure 2). With the model-card pipeline, CT-FM lung risk rose by +1.76 SD (1.56, 2.00) on the thick version of the same scan, ranking agreement fell to 0.34 (0.23, 0.44), and mortality AUC fell by 0.13 (0.05, 0.21). The effect was unchanged in the 147 of the 155 same-kernel pairs with CT-FM embeddings (+1.89 SD; AUC −0.16; −0.30, −0.03), identifying thickness rather than kernel, and persisted after excluding 23 pairs with discordant lung masks (+1.90 SD). The quantitative-CT reference was stable across the same pairs (ranking 0.967; +0.07 SD). Resampling restored stability in a dose-dependent manner (1 mm +0.90 SD; 2 mm +0.06 SD), and the pretraining preprocessing was the most stable setting (−0.02 SD; ranking 0.96). Synthetic 5-mm thickening with fixed coverage reproduced the real-pair values (+1.74 SD; ranking 0.30; AUC −0.12), and changing only the intensity window had no effect (external AUC change −0.006).

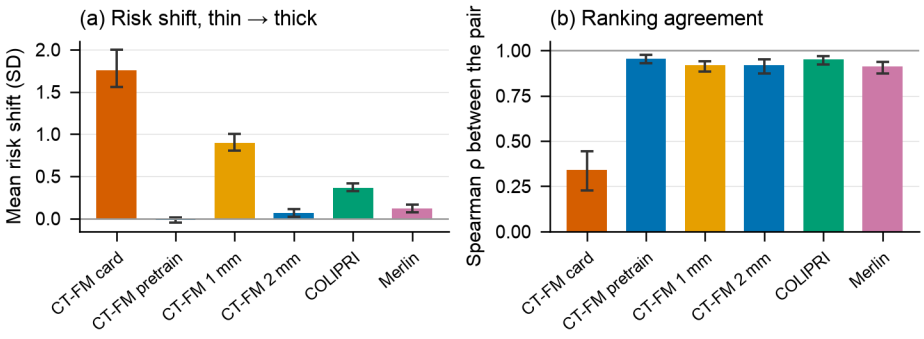


Figure 2: Real thin/thick reconstructions of the same NY-SBU acquisitions (319–333 pairs). (a) Mean shift in predicted risk from the thin to the thick version, in SD of the thin version; (b) agreement of patient ranking between the two versions. "Card" is the model-card pipeline; "pretrain" is CT-FM's pretraining preprocessing (3 × 1 × 1 mm). Whiskers: 95% CI.

Table 2: Real same-acquisition thin (≤ 1.5 mm) and thick (≥ 4.5 mm) reconstruction pairs in NY-SBU (319 pairs for the model-card pipeline, 332 for pretraining, 333 for the others). Risk shift: mean thick – thin logit in SD of the thin logit; ranking: Spearman correlation; AUC change: thick – thin mortality AUC. Lung pooling (Merlin global). Values with 95% CI.

Setting	Risk Shift (SD)	Ranking	AUC Change
CT-FM, model card	+1.76 [1.56, 2.00]	0.34 [0.23, 0.44]	−0.13 [−0.21, −0.05]
CT-FM, pretraining	−0.02 [−0.05, 0.01]	0.96 [0.93, 0.97]	−0.01 [−0.05, 0.02]
CT-FM, 1 mm	+0.90 [0.80, 1.00]	0.92 [0.88, 0.94]	−0.02 [−0.06, 0.01]
CT-FM, 2 mm	+0.06 [0.02, 0.11]	0.92 [0.87, 0.95]	0.00 [−0.03, 0.03]
COLIPRI	+0.36 [0.32, 0.41]	0.95 [0.92, 0.97]	+0.01 [−0.02, 0.03]
Merlin	+0.12 [0.07, 0.16]	0.91 [0.87, 0.94]	+0.02 [−0.02, 0.05]

Reverse interventions produced the same failure in the other models. Removing z-resampling from COLIPRI collapsed its agreement with MosMedData extent grades (Spearman 0.24 to −0.04; difference 0.28; 0.22, 0.34); removing it from Merlin lowered MosMedData agreement (0.32 to 0.24). On MosMedData, 2-mm resampling raised CT-FM lung agreement from −0.04 to 0.15 (+0.19; 0.12, 0.27) and CT-3–4 versus CT-1–2 AUC from 0.52 to 0.70 (Figure 1b shows the agreement values).

Model-Side Repairs Did Not Fully Restore Transfer

Removing site information from the embeddings did not help: a domain-confusion adversary reduced site decodability to chance but never significantly outperformed plain logistic regression, and iterative null-space projection (Ravfogel et al., 2020) produced 2 significant gains and 30 significant losses in 128 comparisons. Fine-tuning the last two COLIPRI blocks matched the frozen probe (+0.015; −0.010, 0.042). Post-hoc device harmonisation removed acquisition coefficients but not the AUC gap,

and thickness augmentation of the training set recovered part of the loss (+0.061; 0.018, 0.104), less than input resampling. External COLIPRI calibration slopes differed by subgroup (0.64 non-contrast, 1.13 contrast; internal 0.94). Intercept-only recalibration on 10–20 labelled target patients captured most of the achievable Brier improvement (0.152 to 0.132–0.134); refitting the slope required 40 or more patients.

Prospectively Specified Test in a Korean Cohort

Across 378 same-acquisition kernel pairs (Figure 3), ranking agreement of the model-card CT-FM pipeline was 0.56 (0.47, 0.63) with the pre-specified probe (age, sex and embedding), compared with 0.78 for the pretraining preprocessing, 0.76 (0.71, 0.80) for 2-mm resampling, 0.66 for Merlin and 0.94 (0.91, 0.95) for COLIPRI. The pre-specified criterion for H-K2 was met (2 mm minus native, +0.20; 0.12, 0.28); an embedding-only probe gave the same conclusion (0.51 versus 0.80). Unlike thickness sensitivity, kernel sensitivity was reduced but not removed by resampling, and a consistent in-plane blur, tested post hoc, did not help. For severity discrimination (240 patients), H-K1 was inconclusive by the pre-specified rule (+0.009; −0.083, 0.126). With near-uniform 3-mm inputs, the thickness shortcut appeared instead as inflated predicted risk in a post-hoc analysis: model-card CT-FM with global pooling predicted a mean probability of 0.36 against a Severe-or-Critical prevalence of 0.17 (all 241 graded patients; Brier 0.179), whereas all other settings predicted 0.08–0.16; intercept recalibration on 10 labelled patients reduced the Brier score to 0.146. In 179 consecutive-visit pairs (28 with a kernel change; embedding-only probe), predicted risk changed more when the kernel changed between visits (model-card CT-FM 0.47 to 0.88 SD; COLIPRI 0.35 to 0.43), an exploratory finding: a protocol change could look like clinical deterioration.

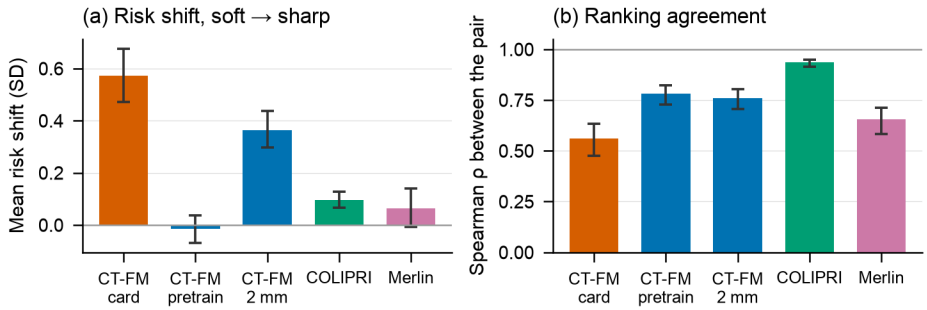


Figure 3: Korean cohort, 378 same-acquisition pairs reconstructed with a soft and a sharp kernel at the same 3-mm thickness: risk shift (a) and ranking agreement (b) for the same settings (pre-specified probe with age, sex and embedding; an embedding-only probe gave 0.51 vs 0.80 for native vs 2 mm, same conclusion). Whiskers: 95% CI.

DISCUSSION

A frozen CT foundation model that appeared to transfer poorly to another hospital was being fed inputs prepared differently from its pretraining.

Following the model card passed each hospital's slice thickness into the embedding, where a probe trained on thin-slice data learned to read thick slices as high risk. The cause was established by intervention, with weights, probe and patients held fixed.

A Latent Condition in the Sociotechnical System

In Reason's (2000) terms, the documentation gap was a latent condition: harmless in the developers' own pipeline, it became a failure only when combined with a site whose scanner protocol differed. Nobody committed an active error, and the usual defences did not catch it. Each individual prediction remained plausible, so a clinician reading a risk score had no cue that it was inflated. In a protocol-uniform cohort, the same shortcut can surface as inflated predicted risk instead, as with global pooling in the Korean cohort. The error is also an equity problem: because it follows acquisition protocol rather than the patient, patients at a site that reconstructs thicker slices tend to receive inflated risk. Errors of this kind are distributed by institution, and could bias any threshold-based triage applied across sites.

Why Fixing the Input Helped More Than Fixing the Model

Attempts to repair transfer by removing site information did not help, consistent with theory under label shift and with domain-generalisation benchmarks (Zhao et al., 2019; Gulrajani and Lopez-Paz, 2021; Koh et al., 2021). Site was decodable from every CT-FM and COLIPRI embedding tested (mean balanced accuracy across training directions 0.82–0.88; chance 0.25–0.33); what distinguished the models was whether an acquisition variable dominated the directions the probe used. Kernel robustness, in contrast, behaved as a model property: COLIPRI was robust and native CT-FM was not; resampling reduced but did not remove this (ranking 0.56 to 0.76, versus 0.94 for COLIPRI), and blurring did not help. The practical lesson for deployment teams is that the cheapest and most effective control sat at the input, before the model, where it is also easiest to audit.

Implications for Model-Card Practice

Model documentation should be treated as a safety-critical interface and versioned and tested like a clinical procedure. For imaging FMs this means, at a minimum, that the model card states the pretraining preprocessing explicitly (orientation, intensity window, crop and voxel spacing), that its usage example executes the same code as pretraining, and that it ships a stability test, such as a paired-reconstruction or synthetic-thickening script with expected values, so that a deploying site can confirm it has reproduced the intended input. Because most of the gap between CT-FM and COLIPRI in our data was preprocessing, and the best FM's advantage over quantitative CT was not established once training-set variability was included, studies that rank FMs should report preprocessing for each model and include training-set variability in their intervals.

A Largely Label-Free Pre-Deployment Check

The findings translate into a short check that a clinical AI team can run before a score reaches a clinician (Table 3). Steps 1 and 2 need no outcome labels; step 3 needs about ten. Same-scan reconstruction pairs are common in hospital archives, because scanners routinely produce thin and thick or soft and sharp reconstructions of one acquisition.

Table 3: Three-step pre-deployment check for a frozen imaging foundation model (steps 1–2 need no labels).

Step	What the Team Does	Signal in this Study
1. Reproduce the pretraining preprocessing	Take spacing, orientation, window and crop from the pretraining code, not the model-card example; if unknown, use a common isotropic spacing.	CT-FM external AUC 0.64 with the model card; 0.71 with pretraining preprocessing; 0.73 at 2 mm.
2. Test stability on same-scan reconstruction pairs (no labels)	Score thin/thick and soft/sharp reconstructions of the same acquisitions from the local archive; compare risk shift and ranking agreement with a reference model.	Ranking 0.34 (thickness) and 0.56 (kernel) for the failing pipeline; 0.91–0.96 for stable settings on thickness pairs.
3. Recalibrate locally (about ten labels)	Refit the intercept on 10–20 labelled local patients; refit the slope only with 40 or more.	Korean cohort Brier 0.179 to 0.146 with 10 patients; NY-SBU 0.152 to 0.132–0.134.

LIMITATIONS

The primary prognostic evidence rests on a single external cohort whose outcome definition differed from that of the training cohort. The public-data hypotheses were generated and then tested in the same cohorts, a circularity we addressed with MosMedData replication, reverse interventions in two other models and a prospectively specified external test, but did not remove. Death was chosen as the primary outcome after the composite outcome had been analysed, and effects were clearest for death and lung pooling and small for the composite outcome and the reverse transfer direction. In the Korean cohort the discrimination hypothesis (H-K1) was inconclusive, and its label was same-time severity rather than prognosis. Several secondary analyses are exploratory. Finally, we have not run a user study: whether clinical AI teams can carry out and interpret the proposed check is untested, and a usability evaluation with such teams is planned.

CONCLUSION

Thickness-driven cross-hospital failure of a frozen CT foundation model arose largely from input preparation that differed from pretraining, and the documentation that users followed was the source of the hazard.

Reproducing the pretraining preprocessing, verifying stability on same-scan reconstruction pairs and recalibrating locally should precede any clinical use or comparison of foundation models at a new site, and model cards should make these steps explicit and testable.

ACKNOWLEDGMENT

This work was supported by Korea Internet & Security Agency (KISA) grant funded by the Korea government (PIPC) (No. RS-2026-25527707, Development of an Interoperable Platform for Pseudonymization and Anonymization of Unstructured Data). This research was also supported by the National Institute of Health (NIH), Korea, research project (project No. 2024-ER0804-02). We thank the providers of the STOIC2021, COVID-19-NY-SBU (The Cancer Imaging Archive) and MosMedData datasets, and the developers of CT-FM, COLIPRI and Merlin for releasing their models.

REFERENCES

Blankemeier, L. et al. (2026). Merlin: A computed tomography vision–language foundation model and dataset. Nature, 652(8112), pp. 1318–1328. doi: 10.1038/s41586-026-10181-8.

Boulogne, L.H. et al. (2024). The STOIC2021 COVID-19 AI challenge: Applying reusable training methodologies to private data. Medical Image Analysis, 97, 103230. doi: 10.1016/j.media.2024.103230.

DeGrave, A.J., Janizek, J.D. and Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence, 3(7), pp. 610–619. doi: 10.1038/s42256-021-00338-7.

Gulrajani, I. and Lopez-Paz, D. (2021). In search of lost domain generalization. International Conference on Learning Representations (ICLR).

Koh, P.W. et al. (2021). WILDS: A benchmark of in-the-wild distribution shifts. Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR 139, pp. 5637–5664.

Lu, L., Ehmke, R.C., Schwartz, L.H. and Zhao, B. (2016). Assessing agreement between radiomic features computed for multiple CT imaging settings. PLoS ONE, 11(12), e0166550. doi: 10.1371/journal.pone.0166550.

Mitchell, M. et al. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), Atlanta, GA, pp. 220–229. doi: 10.1145/3287560.3287596.

Morozov, S.P. et al. (2020). MosMedData: Chest CT scans with COVID-19 related findings dataset. arXiv:2005.06465.

Pai, S. et al. (2025). Vision foundation models for computed tomography. arXiv:2501.09001.

Park, S.H. et al. (2025). Evaluating the robustness of features generated by a foundation model from CT with different reconstruction parameters. Proceedings of SPIE 13411, Medical Imaging 2025: Imaging Informatics. doi: 10.1117/12.3049089.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M. and Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 7237–7256. doi: 10.18653/v1/2020.acl-main.647.

Reason, J. (2000). Human error: Models and management. BMJ, 320(7237), pp. 768–770. doi: 10.1136/bmj.320.7237.768.

Revel, M.-P. et al. (2021). Study of thoracic CT in COVID-19: The STOIC project. Radiology, 301(1), pp. E361–E370. doi: 10.1148/radiol.2021210384.

Roberts, M. et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), pp. 199–217. doi: 10.1038/s42256-021-00307-0.

Saltz, J. et al. (2021). Stony Brook University COVID-19 positive cases [data set]. The Cancer Imaging Archive. doi: 10.7937/TCIA.BBAG-2923.

Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L. and Steyerberg, E.W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. doi: 10.1186/s12916-019-1466-7.

Wald, T. et al. (2025). Comprehensive language-image pre-training for 3D medical image understanding. arXiv:2510.15042.

Zech, J.R. et al. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Medicine, 15(11), e1002683. doi: 10.1371/journal.pmed.1002683.

Zhao, H., Tachet des Combes, R., Zhang, K. and Gordon, G. (2019). On learning invariant representations for domain adaptation. Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97, pp. 7523–7532.