Imperial College London

Department of Computing, Imperial College London

MEng Final Year Project Report

# Fetal Heart Ultrasound Congenital Heart Disease Classification on the iFIND Dataset

**Author:** Abhivir Singh

**Supervisor:** Professor Bernhard Kainz

**Department:** Department of Computing, Imperial College London

**Date:** June 2026

Submitted in partial fulfilment of the requirements
for the degree of Master of Engineering in Computing

# Abstract

Antenatal detection of congenital heart disease on the second-trimester anomaly scan is the largest single screening lever for infant cardiac mortality, yet operates under a severely asymmetric error budget: at the $\sim$1% population prevalence, even a 5% false-positive rate would overwhelm specialist echocardiography services. This dissertation reports a controlled, calibration-aware backbone-swap study on the 4,128-subject iFIND cohort, comparing a Tier 1 baseline built on frozen DINOv2 ViT-B/14 embeddings with a Tier 2 pipeline built on the domain-specialist FetalCLIP ViT-L/14 embeddings (both 768-dimensional) under five identical lightweight aggregation heads, an identical training recipe, five subject-disjoint $k$-fold partitions, and a locked test split. The central result is a dual finding. Tier 2 wins discrimination: the FetalCLIP-backed F2a transformer head reaches an auditable-four AUROC of $0.911 \pm 0.022$ against $0.818 \pm 0.022$ for the DINOv2-backed linear probe E2. Tier 1 wins clinical utility: at the headline screening threshold $p_t = 0.20$, E2's per-fold mean net benefit is $+0.0149$ on five of five folds against F2a's $-0.0372$ on zero of five; the linear probe is the threshold-calibration mechanism. An internal methodology audit walks back a first-pass pooled-on-test-split decision-curve claim and canonises per-fold reporting under $k$-fold cross-validation. Cross-fold temperature scaling narrows but does not robustly close the gap at the per-fold level, leaving the architectural choice — not post-hoc calibration — as the operative mechanism. A bounded scaling experiment localises the discrimination ceiling and falsifies four head-side explanations for it; the asymptotic-budget regime and external validation are the principal directions for future work.

# Acknowledgements

I am grateful to my supervisor, Professor Bernhard Kainz, for sustained guidance throughout this project and for the methodological discipline that shaped its evaluation framework. I thank Dr Tom Day for clinical authority on the fetal-cardiology framing of the screening task, and Matt Baugh for compute infrastructure and engineering support during the scaling-curve window. Thanks also to Lorenzo, Rob, and Sam at Fraiya for project-side collaboration, and to the iFIND team at Imperial College London for the de-identified ultrasound corpus that made this work possible.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 A decision made under a one-per-cent prior

Congenital heart disease is the most common serious birth defect, affecting roughly eight in every thousand live births [1]. For the most severe lesions, the single largest lever on infant survival is not surgical technique but *timing*: a baby whose critical defect is recognised before birth can be delivered in a cardiac centre and stabilised within hours, while one whose defect is missed may collapse at home in the first week of life. In the United Kingdom that lever is pulled, or not pulled, during a few minutes of the routine mid-trimester anomaly scan, when a sonographer sweeps the probe across the fetal chest and decides whether the heart looks normal enough to wave through.

This is a screening decision, and screening decisions live or die by their error budget. The prevalence of congenital heart disease in the general obstetric population is on the order of one per cent, and the cost of the two error types is wildly asymmetric. A missed lethal lesion is catastrophic and, for the conditions that benefit most from antenatal detection, often irreversible. A false alarm, by contrast, is cheap per case but ruinous in aggregate: at a one per cent prior, a test with a five per cent false-positive rate generates roughly five false alarms for every true one, and a national programme that referred at that rate would bury its specialist fetal-echocardiography services within weeks. Any automated tool that hopes to help in this setting must therefore be judged not on how well it separates sick from healthy in the abstract, but on whether it improves the decision a clinician actually makes at a tolerable operating point. That distinction — utility at a usable threshold, not abstract separation — is the spine of this dissertation, and it changes what counts as a good result.

## 1.2 Why the usual metric answers the wrong question

The default yardstick for a binary medical classifier is the area under the receiver-operating-characteristic curve (AUROC). It is a rank statistic: it asks how often the model scores a randomly chosen positive above a randomly chosen negative, averaged over every possible decision threshold. Two properties follow, and both matter here. AUROC is invariant to any monotone rescaling of the scores, so a model can have an excellent AUROC and still be badly miscalibrated — confidently wrong about the absolute risk it reports. And because it integrates over all thresholds, it can reward a model for separating cases at operating points no clinician would ever use, while saying nothing about the one threshold that governs a referral. Two models with identical AUROC can deliver opposite clinical value at the threshold that matters [2].

The appropriate question for a screening tool is therefore not *how well does the model rank?* but *at the threshold a clinician would accept, does acting on the model's output do more good than harm than the alternatives of referring everyone or referring no one?* Decision-curve analysis answers exactly this by reporting net benefit — true positives gained, minus false positives

weighted by the odds of the chosen threshold — as a function of that threshold [3, 4]. A model is clinically useful only where its net-benefit curve sits above both the treat-all and treat-none reference strategies in the threshold band a service would actually adopt. Net benefit depends on calibration in a way that AUROC does not: a model whose scores are systematically inflated will trip the threshold too often and lose net benefit even if its ranking is flawless, and the standard remedy — post-hoc temperature scaling [5] — only helps if the miscalibration is a simple scalar distortion.

This dissertation takes that framing seriously and evaluates every model on both axes at once: discrimination (AUROC, and sensitivity at a fixed high specificity), and clinical utility (per-fold net benefit at a pre-specified screening threshold band). The two axes turn out to disagree, and the disagreement is the result.

## 1.3 The dual finding

The dissertation studies a single, deliberately narrow intervention on the iFIND fetal-cardiac cohort (4,128 subjects, nine congenital conditions, 15.5% research-cohort prevalence): it swaps the frozen image backbone that produces per-frame embeddings, holding the downstream classifier head, the training recipe, the subject-disjoint splits, and the entire evaluation grid fixed. Tier 1 uses DINOv2, a general-purpose self-distilled backbone; Tier 2 uses FetalCLIP, a vision–language model pretrained on fetal ultrasound [6, 7]. Both backbones emit a 768-dimensional per-frame embedding, so the comparison is controlled at the embedding interface, and only the frozen backbone changes. The two image towers are not matched in scale: a smaller ViT-B/14 for DINOv2 against a larger ViT-L/14 for FetalCLIP. The swap therefore varies representation *and* capacity together, and Chapter 7 reports a same-capacity control that separates the two.

The headline result is a tension that I will call the *dual finding*, and it threads through every subsequent chapter. Informally: the domain-specialist backbone wins discrimination, but the simplest head on the general-purpose backbone wins the clinical decision, because the head architecture — not the backbone — is the calibration lever. Stated precisely, as the spine of the chapters that follow:

> Tier 2 (FetalCLIP) wins discrimination; Tier 1 (the DINOv2 linear probe, E2) wins clinical utility; the architectural choice (linear probe) is the calibration mechanism; temperature scaling is the bridge that halves but does not close the gap.

Concretely, on the clinically-grounded auditable-four task (the four UK screening targets, lumped), swapping to FetalCLIP delivers a large, replicated discrimination gain — and yet, at the screening threshold a clinician would actually use, it is the DINOv2 linear probe, not the higher-AUROC FetalCLIP transformer head, whose net benefit clears the treat-none reference on every fold. The better-discriminating model is the less useful one, and the reason is calibration at the threshold: the linear probe has no capacity to inflate mid-range scores, so it keeps its probability mass where a low-prevalence screening decision needs it. A domain-specialist representation, in other words, is necessary neither for the best clinical decision nor sufficient to secure it. Chapters 5 and 6 quantify both halves of this tension.

## 1.4 Research questions

The dissertation is organised around four questions.

**Q1.** Does a domain-specialist vision–language backbone (FetalCLIP) improve *discrimination* on the iFIND congenital-heart-disease task relative to a general-purpose self-distilled backbone (DINOv2), when the downstream head, training recipe, splits, and evaluation grid are all held constant?

**Q2.** Does any discrimination gain transfer into a *clinical net-benefit* gain in the screening operating region ($p_t \in \{0.10, 0.15, 0.20\}$, specificity $\approx 0.95$), or is calibration — rather than discrimination — the binding constraint on clinical utility?

**Q3.** Within a fixed backbone, can the choice of head act as the calibration mechanism; and where it cannot, is a single-parameter temperature scaling [5] enough to bridge the discrimination–utility gap on the screening task?

**Q4.** How does discrimination respond as the per-subject token budget is scaled, and which head-side mechanism — head capacity, training recipe, hierarchical aggregation, or view-aware clip selection — is the binding constraint on the FetalCLIP ceiling?

Questions Q1, Q2, and Q3 are answered directly in Chapters 5 and 6. Question Q4 is the subject of a bounded, pre-registered scaling experiment in Chapter 5: holding the head fixed, we scale the per-subject token budget across several increments and ask, in turn, whether head capacity, training recipe, hierarchical aggregation, or view-aware clip selection is the mechanism that would lift the discrimination ceiling. The asymptotic regime an order of magnitude beyond what we test is left as a sharply-defined open frontier, carried into Chapter 7 with a pre-registered prediction.

## 1.5 Contributions

1. **The dual finding.** A controlled, calibration-aware backbone swap showing that the domain-specialist backbone wins discrimination while the linear probe on the general-purpose backbone wins net benefit, replicated across five subject-disjoint folds, on both the all-nine and the auditable-four tasks, and on both the raw and the postnatal-confirmation-cleaned cohorts. The DINOv2 linear probe is the only configuration whose net benefit clears the treat-none reference on all five folds at the headline threshold.

2. **A per-fold decision-curve methodology, established by self-correction.** An internal audit caught a first-pass scaling result that had compared a pooled-on-test-split net-benefit estimate against a per-fold estimate — an apples-to-oranges comparison that manufactured a spurious win. The dissertation walks the claim back and adopts per-fold net benefit as the canonical estimand for $k$-fold settings [3], and reports the episode as a contribution to how decision-curve analysis should be conducted under cross-validation.

3. **A four-way falsification of head-side scaling hypotheses.** The flat clinical-utility ceiling under token scaling is shown *not* to be a property of head capacity, training-recipe tuning, hierarchical clip aggregation, or cardiac-view-aware clip selection. Of the four, only added head depth moves anything, and it improves aggregate calibration without transferring to the screening threshold — a clean negative result with a mechanistic reading.

4. **A temperature-scaling analysis that qualifies, rather than overturns, the calibration mechanism.** Cross-fold temperature scaling reverses the FetalCLIP-versus-DINOv2 net-benefit gap at the five-fold mean but fails the per-fold replication bar (four of five folds, not five), so the architectural mechanism stands and the fourth clause of the dual finding is qualified, not revised. The fitted temperature sharpens rather than softens, which diagnoses a prevalence-inheritance bias rather than ordinary over-confidence.

5. **A postnatal-confirmation cohort-cleanup procedure.** A filename-encoded confirmation flag in the iFIND release is decoded to restrict the coarctation slice — the condition with the weakest antenatal positive predictive value — to postnatally-confirmed positives, lifting its five-fold AUROC from $\sim$0.85 to 0.879 with no change to any model, and the dual finding survives the cleanup.

## 1.6 Ethical considerations

This work re-analyses a de-identified retrospective research cohort and trains no model intended for deployment; its claims are about evaluation methodology, not about a fielded screening tool. The clinical framing is deliberately conservative throughout: results are read against published sonographer detection rates [8, 9], and the dissertation states plainly where the model falls short of a modern screening programme rather than where it might one day help. The reporting follows the TRIPOD+AI guidance for prediction-model studies [10], item-by-item in Appendix A. Data governance, the single-centre generalisation risk, and the use of generative-AI tooling in preparing this dissertation are set out in full in the Declarations.

## 1.7 How this dissertation is organised

Chapter 2 positions the work against self-supervised and vision–language pretraining, the small set of domain-specialist medical backbones, and the calibration and decision-curve literature, and states the gap the dissertation fills. Chapter 3 describes the iFIND cohort, the nine-condition labelling, the token-budget construction, and the subject-disjoint splits. Chapter 4 establishes the DINOv2 baseline and surfaces the rank inversion — the head that wins discrimination is not the head that wins calibration — that the rest of the argument rests on. Chapter 5 reports the FetalCLIP backbone swap, the discrimination lift, and the bounded scaling experiment behind Q4. Chapter 6 formalises the clinical-utility half of the dual finding through decision-curve analysis, the calibration-mechanism evidence, and the temperature-scaling qualification, and anchors the numbers against a sonographer baseline. Chapter 7 interprets the findings, examines threats to their validity and the alternative explanations, and sets out the open frontier. Chapter 8 concludes.

# Chapter 2

# Background and Related Work

Before the results of this dissertation can mean anything, the reader needs the clinical decision the model is meant to help with, the two visual backbones whose exchange is the whole experiment, and the metrics that decide whether the exchange was worth making — and that is the order this chapter takes them in.

## 2.1 Clinical context: fetal cardiac screening

Congenital heart disease (CHD) is the most common serious birth defect, affecting roughly eight in every thousand live births and rising sharply in tertiary referral populations [1]. It is also, by a wide margin, the group of fetal anomalies most often missed before birth, and the lesions that are missed are disproportionately the ones that kill or disable in the first weeks of life. A defect that is common and lethal yet detectable in principle, and is nonetheless missed in practice, is what makes the heart one of the load-bearing organs of the second-trimester anomaly scan.

In the United Kingdom the scan is delivered by the NHS Fetal Anomaly Screening Programme (FASP), usually between 18 and $20^{+6}$ weeks, as the universal population-level test for structural anomalies [11]. FASP names a fixed set of conditions it is audited against, and four of them are cardiac: hypoplastic left heart syndrome (HLHS), atrioventricular septal defect (AVSD), transposition of the great arteries (TGA) and tetralogy of Fallot (ToF). These four are the targets a national programme is actually accountable for, and they define the 'auditable-four' lumped task that serves as the clinical headline throughout the dissertation (chapter 5, chapter 6).

How the heart is examined is set out in the ISUOG fetal cardiac screening guideline [12], which prescribes a standard sweep through the four-chamber view, the left and right ventricular outflow tracts, and the three-vessel and three-vessel-and-trachea views, with situs and rhythm checks along the way. In expert hands the combined four-chamber-plus-outflow examination reaches a sensitivity near 89% at a specificity close to 99.98%, but that figure is a ceiling, not a floor. Routine detection outside specialist centres is far lower and uneven across conditions: pooled estimates for prenatal detection of major CHD have historically sat below half of cases [8], and even recent UK audit ranges span roughly 69% to 93% depending on lesion and centre [9]. A model that hopes to help has to be read against that real spread, not against the expert ceiling.

The decisive point for everything that follows is that screening is statistically a different problem from diagnosis. A fetus flagged at the 20-week scan is referred onward to fetal echocardiography, where a cardiologist performs a longer, definitive examination. The screening step itself sits at a very-high-specificity operating point because the prior is so low: with CHD prevalence on the order of one per cent in an unselected cohort, a test that called even 5% of healthy fetuses abnormal would produce roughly five false referrals for every true one and swamp the specialist services it feeds. The dissertation is explicit about living in the screening regime. Its pre-specified operating points are specificity 0.95 and specificity 0.99. The lower point is a deliberately sub-clinical threshold, chosen because the cohort cannot resolve sensitivity

differences at 0.99 with adequate precision. Both feed a decision-curve threshold-probability grid $p_t \in \{0.10, 0.15, 0.20\}$ read off the same screening framing, whose meaning is unpacked in Section 2.5.

## 2.2 The iFIND corpus

The cohort comes from iFIND (intelligent Fetal Imaging and Diagnosis), the Imperial-led, EPSRC- and Wellcome-funded programme that assembled a large de-identified corpus of fetal ultrasound with the stated aim of changing how the 20-week anomaly scan is done. The same clinical and engineering team went on to run PROMETHEUS, a randomised controlled trial of AI-assisted fetal anomaly screening that takes CHD detection as its primary outcome precisely because CHD is the most missed and most mortality-relevant of the screened conditions [13]; PROMETHEUS is the most recent peer-reviewed anchor for the data pipeline this dissertation builds on.

The slice used here is 4,128 subjects carrying nine condition labels. The nine-condition schema, the subject-disjoint 2,888/620/620 train/validation/test split, the construction of the auditable-four lumped binary task, and the postnatal-confirmation cleanup of the coarctation-of-the-aorta column are all set out in chapter 3; this chapter only needs the reader to know that the cohort is real, single-programme, and of a size that constrains how finely per-condition performance can be resolved.

## 2.3 Vision foundation models for medical imaging

### 2.3.1 Vision transformers and self-supervised pretraining

The Transformer [14] replaced recurrence and convolution with multi-head self-attention — a primitive in which every token attends to every other and the model learns which relationships matter — and went on to become the default backbone across language and, eventually, vision. The Vision Transformer (ViT) [15] carried the idea to images by cutting a picture into fixed-size patches, projecting each patch into a token embedding, and running the resulting sequence through a standard encoder stack. Given enough pretraining data, this pure-attention design matched or beat convolutional networks on image recognition while training more efficiently.

The way these backbones are pretrained has shifted at least as much as the architecture. Supervised ImageNet pretraining of convolutional networks dominated medical-imaging transfer until around 2020, but its restricted label vocabulary and photographic-domain bias capped how well it carried to greyscale clinical imagery. Large-scale *self-supervised* pretraining — on unlabelled corpora orders of magnitude larger than ImageNet — has since become the default source of general-purpose features. Masked image modelling, instance-discrimination contrastive learning, and self-distillation are the objective families that dominate this space, and the dissertation builds on the last of them.

### 2.3.2 DINOv2: self-distillation at scale

DINO [16] cast self-supervised pretraining of vision transformers as self-distillation with no labels. Two augmented crops of one image go to a student and a teacher network; the student is trained to match the teacher's output distribution, while the teacher is an exponential moving average of the student. The features that fell out of this loop turned out to encode explicit segmentation structure and to act as strong nearest-neighbour classifiers with no fine-tuning at all.

DINOv2 [6] scaled that recipe along three axes: a curated 142-million-image corpus, larger backbones up to ViT-g/14, and a stabilised loss and optimiser. A family of distilled checkpoints was released alongside the largest model. The one this dissertation uses is the distilled ViT-B/14 (`facebook/dinov2-base`), which emits a 768-dimensional patch-level embedding learned without any labels. The authors show these features transfer to a broad range of downstream tasks,

including several medical-imaging benchmarks, without supervised fine-tuning, which is the warrant for treating DINOv2 as a frozen feature extractor. The Tier 1 baseline (chapter 4) does exactly that: per-frame embeddings are read off the frozen ViT-B/14 and a small head — a linear probe or a shallow transformer — is trained on top.

### 2.3.3 CLIP-style vision–language pretraining

CLIP [17] swapped the supervised classification target for a contrastive image–text one. An image encoder and a text encoder are trained together on 400 million web image–caption pairs so that an image embedding sits close to the embedding of its own caption and far from mismatched ones. The payoff is a shared image–text space in which classification can be done at inference time by encoding candidate label prompts and taking the nearest match, with no task-specific head trained at all. CLIP set the contrastive image–text recipe as the dominant route to *vision–language* foundation models.

CLIP and DINOv2 are best contrasted by *what kind of pretraining signal each uses*, not by how much supervision each gets. Neither uses task-specific human-labelled class supervision: DINOv2 is image-only and self-supervised, CLIP uses paired image–text but no diagnostic labels. The difference is the alignment target. DINOv2 builds features invariant to image augmentations; CLIP builds features whose neighbours in text space are semantically meaningful captions. The two recipes therefore carve out qualitatively different feature geometries. The dissertation isolates that difference cleanly by feeding both tiers' per-frame embeddings — each 768-dimensional — into the same set of subject-aggregation heads under the same training recipe, so that any Tier 1-versus-Tier 2 gap traces to the backbone (its pretraining objective and its image-tower scale) rather than to the downstream model.

### 2.3.4 FetalCLIP: a domain-specialist CLIP for fetal ultrasound

FetalCLIP [7] is the domain-specialist CLIP used as the Tier 2 backbone. Its image tower is a ViT-L/14 (24 layers, width 1024, patch size 14) initialised from a published CLIP checkpoint and fine-tuned jointly with the text tower on 210,035 fetal-ultrasound image–caption pairs. Although the tower is larger than the Tier 1 ViT-B/14, FetalCLIP projects its output to a 768-dimensional embedding — the same width DINOv2 emits — so the two backbones present an identical interface to the heads downstream. That matched width is what makes the backbone swap of chapter 5 a like-for-like exchange rather than a confounded change of input dimensionality, with image-tower scale (ViT-B versus ViT-L) named as the one capacity confound the swap deliberately carries.

The captions are not free-text clinician dictations. They were produced by templating a large language model over structured clinician annotations — view labels such as 'four-chamber view', gestational age, pixel-spacing metadata — into synthetic but anatomically faithful frame descriptions. The corpus splits into two strata: the bulk, drawn from routine second-trimester scans that carried no information on fetal health status, and a small stratum of roughly one per cent sourced from a fetal ultrasound textbook and expert annotators that emphasises critical cardiac conditions. That thin cardiac stratum holds the only explicit disease-concept supervision the pretraining ever sees.

This matters for how the Tier 2 discrimination win should be read. No part of FetalCLIP's pretraining objective is disease-supervised in the conventional sense: the contrastive loss matches an image to a caption describing anatomy and metadata, never to a diagnostic label for the CHD task this dissertation studies. The uplift FetalCLIP delivers on iFIND therefore reflects representation quality earned from clinician-grounded captions, not leaked diagnostic labels — which is exactly what lets the backbone swap support a clean architecture-effect interpretation. chapter 7 returns to this point.

## 2.4 Medical foundation models: what does and does not transfer

FetalCLIP is one member of a fast-growing family of medical foundation models that followed the generalist-AI turn in medical imaging argued for by Moor and colleagues [18], who describe a move from a model-per-task world to one in which a few large multimodal models, pretrained on heterogeneous data, are adapted to specific clinical tasks with little task-specific labelling. That programme depends on the self-supervised and contrastive recipes of the previous section, but it demands domain-specific evaluation, because medical imagery breaks many assumptions of natural-image pretraining: it is greyscale and narrow-field, its texture is dominated by speckle, and its appearance shifts with the scanner that produced it. The controlled comparison this dissertation runs is one concrete instance of that evaluation question — when a domain-specialist backbone shares the image-tower lineage of a general-purpose one, does it win for the right reason?

The neighbours below are the obvious comparators, and each is instructive less for what it shares with FetalCLIP than for what fails to carry across to fetal cardiology.

**BiomedCLIP.** BiomedCLIP [19] is a CLIP-style biomedical model trained on PMC-15M, fifteen million image–caption pairs scraped from the figures and captions of PubMed Central articles, paired with a PubMedBERT text encoder. It is the natural CLIP-on-biomedical-data contrast to FetalCLIP, but its corpus lives at the *literature-figure* level: modality- heterogeneous, dominated by static publication figures and microscopy, and processed for print rather than captured at the scanner. What does not transfer is precisely the thing fetal cardiology turns on — temporal, in-vivo, single-modality sweep video. A model whose visual prior was set by composited journal figures has seen almost nothing of a beating fetal heart.

**CONCH.** CONCH [20] is a vision–language pathology model trained on 1.17 million histopathology image–caption pairs, and its design mirrors FetalCLIP closely: a contrastive image–text loss, a carefully curated captioning pipeline, a ViT image tower, with strong results across many computational-pathology tasks. The lesson it carries is encouraging — in-domain CLIP pretraining, given enough pairs, can underwrite a wide downstream range. The limit is just as clear. Histopathology is static, high-magnification tissue imaged ex vivo, with no motion, no acquisition sweep and no cardiac dynamics; its representation prior says nothing about which of twenty noisy ultrasound frames of a moving heart is even diagnostic. CONCH shows the recipe generalises within a domain, not across to one defined by motion.

**USFM.** USFM [21] is the closest non-fetal alternative: a universal ultrasound foundation model self-supervised on more than two million ultrasound images across cardiac, abdominal, obstetric and musculoskeletal organs, many centres and many devices, using a spatial-frequency masked-image-modelling objective built for speckle. It demonstrates that ultrasound supports its own scaling laws. Two things still do not transfer to the present task. USFM is image-only and so forgoes the text signal that FetalCLIP's captions inject, and it is general rather than fetal-cardiac, so its prior is spread across organs whose appearance statistics differ sharply from a 20-week heart. FetalCLIP's bet is that the fetal sub-domain is distinctive enough to repay a specialist-corpus-with-text recipe over a universal-ultrasound-image-only one. The dissertation does not benchmark USFM directly, and flags it in chapter 7 as the natural next comparison.

**Why fetal cardiac ultrasound resists all of them.** What makes none of these priors transfer cleanly, and a fetal specialist plausibly worth the trouble, comes down to the task itself. The field of view is small and motion-rich: the heart at 20 weeks is about two centimetres across, beating near 140 beats per minute, with the probe in constant motion, so most individual frames are non-diagnostic. Doppler and B-mode frames share a single clip stream yet carry very different

Table 2.1: Where FetalCLIP sits among contemporary medical foundation models, and what fails to carry across to fetal cardiac ultrasound. 'In-domain' means fetal-ultrasound-specific pretraining.

| Model | Pretraining signal | Domain | Motion? | Gap for fetal cardiology |
|---|---|---|---|---|
| BiomedCLIP | image–text | biomedical literature | no | static figures, mixed modality |
| CONCH | image–text | histopathology | no | ex-vivo tissue, no cardiac dynamics |
| USFM | image-only (MIM) | general ultrasound | frames | no text signal, not fetal-specific |
| FetalCLIP | image–text | fetal ultrasound | frames | (in-domain; not disease-supervised) |

appearance statistics, opening the door to shortcut features that chapter 3 treats explicitly. And the view convention is hierarchical and disease-specific — HLHS is readable on the four-chamber view alone, TGA needs the outflow tracts — a structure a view-agnostic embedding pipeline may under-exploit. These are exactly the conditions a literature-figure, a histopathology slide, or a general-ultrasound prior was never shaped against.

Table 2.1 gathers the comparison.

## 2.5 Decision-curve analysis and net benefit

The default discrimination metric for a binary clinical classifier is the area under the receiver-operating-characteristic curve (AUROC), and two of its properties make it a poor *primary* metric for a screening test, as Halligan, Altman and Mallett set out [2]. AUROC is invariant under any monotone rescaling of the scores, so two classifiers with identical AUROC can disagree completely about how well-calibrated they are at a fixed threshold. And it weights the whole ROC space uniformly, including operating points no clinician would deploy, so for a test that must run at very high specificity the headline number can be dominated by behaviour far from the threshold that actually governs a referral.

Decision-curve analysis (DCA), introduced by Vickers and Elkin [3], addresses both limits head on. For a chosen threshold probability $p_t \in [0, 1]$ it reports the model's *net benefit*,

$$\mathrm{NB}(p_t) = \frac{\mathrm{TP}(p_t)}{n} - \frac{\mathrm{FP}(p_t)}{n} \cdot \frac{p_t}{1 - p_t},$$

where $p_t$ is the cost trade-off the analyst is willing to assert: intervene on a subject whose predicted probability is at least $p_t$. The weight $p_t/(1 - p_t)$ is the relative cost of a false positive against a missed case; at $p_t = 0.20$ it says one false-positive referral is worth one-quarter of a missed lesion. A model earns its place only where its net benefit exceeds both default strategies — treat-none, with $\mathrm{NB} = 0$, and treat-all, whose net benefit falls with $p_t$ and prevalence — across the threshold band a service would actually adopt. Below either default, the model is worse than deciding without it.

DCA has since become a standard adjunct to discrimination and calibration reporting in oncology, cardiology and prediction research, helped by the step-by-step interpretive guide of Vickers, Van Calster and Steyerberg [4] and by editorial endorsement across the major clinical journals. The dissertation uses it as the primary clinical-utility metric throughout chapter 6, on the screening grid $p_t \in \{0.10, 0.15, 0.20\}$ with $p_t = 0.20$ as the headline.

DCA here is computed *per fold* on the subject-disjoint five-fold partition, never pooled across a single fixed test split — a commitment that shapes every utility number in the report. Pooling is biased when the per-fold models are trained on different subsets and calibrated on different validation sets, and chapter 5 treats one concrete instance of that bias explicitly. The per-fold convention takes its discipline straight from the Vickers and Elkin prescription for evaluating prediction models on independent samples [3].

## 2.6 Calibration, reliability, and probabilistic forecasting

A predicted probability $\hat{p}(x)$ is *calibrated* if, among all cases assigned probability $q$, the observed positive rate really is $q$. Calibration is what turns a score into a usable risk estimate rather than a bare ranking, and it is the property net benefit silently reads off at whatever threshold the clinician sets — which is why a chapter on discrimination cannot stop at AUROC. The dissertation pins calibration down with three instruments, each doing a job the others cannot: Murphy's decomposition of the Brier score [22] splits one scalar into a calibration term and a discrimination term, Guo and colleagues' temperature scaling [5] tests whether a single parameter can repair the calibration without touching the ranking, and Efron's BCa bootstrap [23] keeps the intervals honest where the threshold metrics turn non-smooth.

**The Brier score and its Murphy decomposition.** The Brier score is the mean squared error between predicted probability and binary outcome, $\text{Brier} = \frac{1}{n}\sum_i(\hat{p}_i - y_i)^2$. Murphy showed it decomposes exactly into three additive parts [22],

$$\text{Brier} = \text{REL} - \text{RES} + \text{UNC},$$

where reliability (REL) measures how far binned predictions drift from binned empirical frequency, so lower is better calibrated. Resolution (RES) measures how far those frequencies move away from the base rate, so higher is more discriminating. Uncertainty (UNC) is fixed by the base rate alone and owes nothing to the classifier. The decomposition factors one scalar into a calibration term and a discrimination term, and underpins the Brier–Murphy analysis later in the dissertation. For the auditable-four task its prevalence is roughly 0.10, which pins UNC near 0.093; for the all-nine task UNC is near 0.131.

Two cautions follow, and both turn out to be load-bearing for the dual finding. REL is *aggregate* calibration, averaged across the whole $[0, 1]$ interval, so a model can look well-calibrated globally and still be badly off in the narrow neighbourhood of the screening threshold $p_t$. Two models with identical average calibration can diverge entirely at a screening threshold of 0.20 if one parks its probability mass just below the threshold and the other just above: their REL scores agree, their referral decisions do not. This distinction between aggregate and threshold calibration, made precise by Van Calster and colleagues [24], is exactly the mechanism chapter 6 uses to reconcile an aggregate-calibration loser winning the decision-curve contest — the conceptual hinge on which the whole utility result swings.

**Temperature scaling.** Deep classifiers are characteristically over-confident, and Guo and colleagues show that much of that over-confidence is a single scalar distortion: divide the pre-sigmoid logits by one learned temperature $T$, fitted by minimising negative log-likelihood on held-out data, and the reliability curve often snaps back into place [5]. The appeal for this work is precisely that $T$ is a monotone rescaling — it leaves the ranks, and therefore AUROC, untouched, so any change it produces is calibration and nothing else. That makes it the sharpest possible test of one question the dual finding turns on: chapter 6 applies it to ask whether the FetalCLIP-versus-DINOv2 calibration gap is the kind of simple distortion a single scalar can erase, or something the architecture has built in.

**BCa-bootstrap confidence intervals.** The headline metrics here are discrete-threshold quantities — sensitivity at fixed specificity, PPV at fixed specificity, net benefit at a fixed $p_t$ — and that is exactly what makes them awkward to put an interval around. Each is a non-smooth function of the prediction distribution with skewed sampling behaviour, and the skew bites hardest at the small positive counts of the per-condition slices, the regime where the ordinary percentile bootstrap quietly runs over-conservative and overstates the precision on offer. Efron's bias-corrected and accelerated (BCa) bootstrap [23] corrects both the median bias and

the skewness, recovering nominal coverage at a tighter interval, which is why it is the interval engine behind every per-condition AUROC, PPV and net-benefit estimate in Chapters 4–6, run subject-level with 1,000 resamples.

## 2.7 Clinical-prediction-model reporting standards

Reporting discipline is not decoration in a prediction-model study; it is what lets a reader tell a real result from a favourably-sliced one. The TRIPOD statement (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) codifies that discipline as a checklist, and its 2024 extension by Collins and colleagues [10], TRIPOD+AI, extends the 27-item guidance to cover machine-learning as well as regression methods, from title and abstract through methods, open science, results and discussion. The dissertation reports against it item by item, with the full table in the appendix and a compliance summary in chapter 7. Two methodological commitments made elsewhere in the report are not housekeeping but direct answers to specific TRIPOD+AI items: per-fold DCA and BCa intervals are how the work meets the demand for performance estimates carrying proper uncertainty, and the item on training-versus-evaluation data is what forces the clean line between the internally-validated models produced here and the external validation deferred to future work.

Two further standards set the bar against which two of the work's choices are read. The Riley sample-size calculation for clinical prediction models [25] normally sizes a cohort before collection; with iFIND fixed, the dissertation inverts it, applying the calculation after the fact to quantify the per-condition power budget in chapter 3 — and so to say plainly where the cohort is too thin to resolve a difference. Calibration reporting, in turn, is graded against the Van Calster hierarchy — mean, weak, moderate, strong [24] — and the auditable-four analysis in chapter 6 is held to the moderate level rather than claimed at a level the data cannot support.

## 2.8 The gap

Everything above narrows to one absence. The medical foundation-model literature reports domain-specialist backbones that win discrimination, but it almost never asks whether that win survives contact with the clinical decision the model is meant to support, and it rarely isolates the backbone from the head that sits on top of it. The FetalCLIP, USFM and BiomedCLIP papers each report discrimination-only or single-task evaluations, on cohorts and pipelines that confound the representation with the classifier and the evaluation with the operating point.

What no prior fetal-cardiac study does is the controlled, calibration-aware experiment this dissertation runs: swap a general-purpose self-distilled backbone for a domain-specialist vision–language one at a *matched* 768-dimensional interface, holding the head, recipe, splits and metric grid fixed, on a real cohort of roughly four thousand subjects, and grade the result *jointly* — on discrimination (AUROC and sensitivity at specificity 0.95) and on per-fold decision-curve net benefit at a pre-specified screening threshold band. Reading those two axes against each other, rather than treating the first as a proxy for the second, is what produces the dual finding of chapter 6. The next chapter describes the iFIND cohort and its labelling in the detail the methods and results require.

# Chapter 3

# Data, Labels, and Splits

Every claim in the chapters that follow rests on a single retrospective cohort, a labelling scheme that began life as a clinical report rather than a machine-learning target, and a partitioning discipline designed to keep the same fetus from appearing on both sides of an evaluation. This chapter sets those foundations out in enough detail to reproduce, and it is candid about where the labels are softer than the headline numbers might suggest. Two decisions in particular shape the reading of everything downstream: the choice to report a clinically-grounded four-condition headline alongside the broader nine-condition derivation, and a targeted cohort cleanup on the one condition whose antenatal label is known to be unreliable.

## 3.1 The iFIND cohort

The empirical work draws entirely on the iFIND research cohort, a retrospective collection of second-trimester fetal ultrasound video captured during routine ∼20-week anomaly scans at Imperial-affiliated hospitals as part of the iFIND fetal imaging programme introduced in Chapter 2. The released cohort holds 4,128 unique subjects, each represented by a variable number of short clips from a single antenatal visit. For the experiments of Chapters 4–6 we admit up to twenty clips per subject and ten frames per clip, a canonical per-subject budget of 200 frames; the construction of that budget is the subject of Section 3.6. All clips are de-identified at source under the iFIND project's institutional ethics and data-sharing approvals, and the dissertation collects no new human-participant data. The ethical framing introduced in Section 1.6 is extended in Chapter 7.

Of the 4,128 subjects, 640 carry at least one of the nine congenital heart disease labels described in Section 3.2, an any-CHD prevalence of 15.50%. The remaining 3,488 subjects (84.50%) carry no positive label and form the healthy reference class throughout, subject to the caveats of Section 3.8. The positive class is not a tidy partition of single diagnoses: among the 640 unhealthy subjects, 533 carry exactly one positive label, 103 carry two, and 4 carry three or more, so 107 of 640 (16.7%) present at least one comorbid CHD diagnosis. That overlap is why the per-condition decomposition of Section 3.5 keeps comorbid subjects in every positive set they qualify for rather than forcing a single primary diagnosis.

## 3.2 The nine-condition labelling schema

The subject-level label table carries nine binary condition columns, each populated by the iFIND clinical team from the antenatal scan reports. Table 3.1 lists the nine conditions with their full names and prevalence in the released cohort. The columns are independent 0/1 indicators in the data-modelling sense, but they are not clinically exclusive: a subject with hypoplastic left-heart syndrome and a coexisting coarctation is positive on two columns. The any-CHD label used throughout is the disjunction across all nine.

| column | condition | $n_{\text{pos}}$ (full cohort) | prevalence |
|---|---|---|---|
| tga | transposition of the great arteries | 128 | 3.10% |
| tetralogy | Tetralogy of Fallot | 115 | 2.79% |
| raa | right aortic arch | 109 | 2.64% |
| coa | coarctation of the aorta | 101 | 2.45% |
| avsd | atrioventricular septal defect | 86 | 2.08% |
| hlhs | hypoplastic left-heart syndrome | 83 | 2.01% |
| p_atresia | pulmonary atresia | 74 | 1.79% |
| p_stenosis | pulmonary stenosis | 28 | 0.68% |
| a_stenosis | aortic stenosis | 27 | 0.65% |
| unhealthy | OR-of-9 (derived) | 640 | 15.50% |

Table 3.1: The nine subject-level CHD condition columns and the derived any-CHD label, with positive counts and prevalence on the full 4,128-subject iFIND cohort. The columns are independent binary indicators; the OR-of-nine is computed downstream and is not a stored field in the canonical labels file.

The nine columns are themselves a coarsening. They were derived from a finer 43-category clinical subdivision through a fixed $43 \to 9$ lookup table maintained by the project's clinical-informatics collaborator.[1] The empirical work uses the nine-column labels as supplied; the finer 43-way subdivision appears in Chapters 5–6 only as a within-nine error-analysis pointer, never as a training target.

One property of the schema bears directly on the cleanup that follows. The labels are read from *antenatal* scan reports, so they are antenatal-suspicion labels rather than postnatal-confirmation labels. For eight of the nine columns antenatal diagnosis approximates ground truth closely enough to serve as a training target.[2] The exception is coarctation of the aorta, which is well documented as one of the hardest cardiac lesions to call before birth: the positive predictive value of an antenatal CoA suspicion is roughly 0.5, meaning that about half of the subjects flagged `coa=1` on the antenatal scan are not confirmed to have CoA at birth.[3] The targeted cleanup that follows from this is described in Section 3.4.

A second, subtler property runs through the whole study and is taken up in full in Section 3.8: a zero in any condition column means the subject does not carry that antenatal label, not that the subject was examined and ruled out, so the healthy class is a not-labelled class rather than a verified-negative one.

## 3.3 The auditable-four task: the UK FASP screening headline

The clinical headline of the dissertation is reported on a subgroup, not on the full nine-condition derivation, and the reason is comparability. The UK Fetal Anomaly Screening Programme grades its services on a defined set of cardiac targets, and a model whose performance is reported on a different subgroup cannot be read against the audited performance of the existing service. The FASP handbook [11] names four cardiac conditions as the screening targets assessed at the routine ∼20-week scan: transposition of the great arteries, atrioventricular septal defect, Tetralogy of Fallot, and hypoplastic left-heart syndrome. These four anchor the international cardiac-screening guidance as well, in which the four-chamber and outflow-tract views are the primary objects of mid-trimester cardiac assessment [12].

We therefore define the *auditable-four* task: a lumped binary in which the positive class is

[1] Fraiya project, personal communication, 2024.

[2] Consultant fetal cardiologist, personal communication, 2026.

[3] Clinical collaborator, personal communication, consistent with the published literature on antenatal CoA.

the set of subjects with at least one of {TGA, AVSD, Tetralogy of Fallot, HLHS}. Co-occurring labels on other conditions are disregarded for membership, so a subject positive on both AVSD and CoA is an auditable-four positive. The negative class is the healthy reference class — the 3,488 subjects with zero on every condition column — not the complement of the positive class. This healthy-only-negative convention sharpens the clinical question: the model is being asked whether it can separate the four flagship lesions from the screening population, not from other cardiac pathology. A sensitivity check that widens the negative class to include the non-auditable CHD subjects (RAA, CoA, pulmonary atresia, aortic and pulmonary stenosis) is reported in Section 5.4; it shifts every AUROC downward by roughly 0.02 without disturbing the architecture ranking.

The auditable-four positive class is a subset of the any-CHD positive class sharing the same negatives, so its prevalence is lower. Across the five subject-disjoint folds the auditable-four prevalence is approximately 0.104 (around 81 positives in roughly 779 subjects per fold, about 698 healthy negatives), against the all-nine prevalence of 15.50%. The auditable-four lumping is the primary clinical headline of Chapters 5 and 6; the all-nine binary is the secondary report. Reporting both protects against a subgroup chosen to flatter the result, and the dual finding of Chapter 6 holds on both.

## 3.4 Postnatal-confirmation gating and the CoA cleanup

Coarctation is the condition where the antenatal label is least trustworthy, and it is worth fixing rather than carrying the noise into every per-condition number. The canonical labels file holds no postnatal-confirmation column, but a confirmation indicator turned out to be encoded in the per-clip filename metadata as a 0/1/2 'yes/no/unknown' field, on the clinical collaborator's advice that it be applied only to the CoA slice. Each filename basename across all 197,088 released clips is composed of exactly ten underscore-separated tokens, which we index `tok[0]`–`tok[9]` in order of appearance. The confirmation field was not documented in the data release, so its position and value mapping were recovered empirically. Recovering it is an originality contribution rather than a housekeeping detour: it is the lever that converts the cohort's least trustworthy label into a postnatally-gated one, and the per-condition CoA gain reported in Section 5.6 is downstream of the decoding work that follows.

Two of the ten tokens carry value-set cardinalities compatible with a 0/1/2 field. `tok[4]` takes five values and tracks the healthy/unhealthy split, consistent with a view-or-quality code; `tok[5]` takes three values plus a 48-frame anomaly, ranging over $\{0, 1, 2, 3\}$. The mapping was pinned down by cross-referencing `tok[5]` against the `tok[3]` free-text sonographer diagnosis. Of the 101 subjects with `coa` $= 1$, the 23 at `tok[5]`=0 carry zero CoA mentions in `tok[3]`. A curator would not retain CoA free-text on a subject whose coarctation was ruled out at birth, so this aligns with postnatal-not-confirmed. The 36 at `tok[5]`=1, by contrast, show a 67% CoA-mention rate, consistent with postnatal-confirmed-positive. The canonical mapping used in code and throughout is therefore {0: no, 1: yes, 2: unknown, 3: unknown} on `tok[5]`. Two checks corroborate it: healthy subjects sit overwhelmingly at `tok[5]`=2 (3,481 of 3,488, since no postnatal check is needed for normals), and non-CoA unhealthy subjects sit at `tok[5]` $\in \{1, 2\}$ with no postnatal-no calls.

The gate is applied only to the CoA slice. A CoA-1 subject is retained if either `tok[5]`=1 (postnatally confirmed) or the subject is positive on at least one of the other eight columns, in which case the comorbid condition anchors the subject as unhealthy regardless of CoA status. A single-labelled CoA-1 subject with `tok[5]` $\in \{0, 2, 3\}$ is dropped from training, validation, and test rather than relabelled healthy, on the collaborator's advice that these subjects 'don't really fit normal or not-normal'.

Table 3.2 reports the cohort delta. The full cohort shrinks from 4,128 to 4,086 subjects, a drop of 42, all of them single-labelled CoA suspicions. The CoA-positive count falls from 101

to 59: 36 postnatally confirmed plus 23 anchored by a comorbid non-CoA condition. All eight non-CoA condition columns keep their original counts in every partition, because the rule can only ever drop subjects with `coa = 1` and zero on every other column.

| partition | $n$ before | $n$ after | dropped | CoA-1 before | CoA-1 after |
|---|---|---|---|---|---|
| test | 620 | 617 | 3 | 14 | 11 |
| val | 620 | 616 | 4 | 14 | 10 |
| train | 2888 | 2853 | 35 | 73 | 38 |
| full | 4128 | 4086 | 42 | 101 | 59 |

Table 3.2: Cohort delta from the postnatal-confirmation cleanup on the CoA slice. The cleanup drops 42 single-labelled CoA-suspicion subjects who lack postnatal confirmation; the 59 surviving CoA-1 subjects are 36 postnatally confirmed plus 23 anchored by a comorbid non-CoA condition. All non-CoA condition counts are unchanged in every partition.

The payoff is reported in Section 5.6. On the cleaned cohort the FetalCLIP per-condition CoA AUROC rises to $0.879 \pm 0.023$ across five folds, against an approximate baseline of 0.85 on the un-cleaned antenatal-suspicion cohort: a gain of about +0.03 AUROC with no change to any model, attributable entirely to labelling discipline. The auditable-four headline is unaffected ($0.911 \pm 0.022$ on both cohorts), since CoA is not one of the four. The cleanup sharpens the training and evaluation signal, but it cannot manufacture a stronger antenatal sign than exists. Coarctation's antenatal positive predictive value is inherently weak, and a model trained and tested on a postnatally-cleaned slice is still being asked to detect a lesion that fetal cardiologists themselves call late; the improved AUROC reflects cleaner labels, not a claim that antenatal CoA detection is a solved problem.

## 3.5 Per-condition slicing-$i$ decomposition

For each of the nine columns we additionally define a *slicing-i* binary task: the positive class is the set of subjects positive on the column under analysis, regardless of comorbidity, and the negative class is the healthy reference class. This is one-versus-rest restricted to a clean healthy-only negative set, and it serves throughout Chapters 4–6 as the per-condition decomposition behind the lumped headlines. The healthy-only-negative convention keeps non-target CHD signal out of the negatives, and the comorbidity-preserving positive convention matches the rule used to derive the all-nine binary.

Per-condition positive counts on the held-out 620-subject stratified test split appear in Table 3.3, alongside the across-fold mean for the subject-disjoint five-fold replication that produces the headline intervals.

Sample size, not method, sets the ceiling on any per-condition claim. AVSD, TGA, Tetralogy, HLHS, RAA, CoA, and pulmonary atresia each contribute roughly eight to twenty-five positives per fold; aortic and pulmonary stenosis each contribute fewer than six. The load-bearing inferential claims are made on the lumped auditable-four task (around 78–84 positives per fold) and on the all-nine binary (roughly 128 positives per fold), not on per-condition rows. The per-condition decomposition is descriptive context, and Chapter 5 treats it as such.

## 3.6 Token budget, training, validation, and test splits

The per-subject input to every model in this dissertation is a fixed token budget, and how that budget is built matters as much as how large it is. Each subject contributes up to twenty clips, and from each clip we sample ten frames at indices spaced evenly across the clip's full duration

| condition | $n_{\text{pos}}$ (test split) | $n_{\text{pos}}$ (5-fold mean) |
|---|---|---|
| TGA | 21 | 25.6 |
| Tetralogy of Fallot | 16 | 23.0 |
| RAA | 13 | 21.8 |
| AVSD | 17 | 17.2 |
| HLHS | 10 | 16.6 |
| Pulmonary atresia | 14 | 14.8 |
| CoA (cleaned) | 11 | 11.8 |
| Pulmonary stenosis | 2 | 5.6 |
| Aortic stenosis | 5 | 5.4 |

Table 3.3: Per-condition positive counts under the slicing-$i$ decomposition. The single-split column reports the held-out 620-subject stratified test split; the 5-fold column reports the average per-fold positive count across the five subject-disjoint folds. The CoA row uses the cleaned cohort of Section 3.4. The bottom two rows (pulmonary stenosis, aortic stenosis) have $n_{\text{pos}} < 6$ on the test split and are reported in Chapter 5 as exploratory; see Section 3.8.

rather than as a contiguous run — a `linspace` over the clip length. The intent is structural rather than kinematic: ten evenly-spaced frames sample the heart across the phases of the cardiac cycle, which is what a frozen image encoder can use, where ten consecutive frames would capture a fraction of a single beat. With ten frames over twenty clips, the canonical budget is 200 frames per subject, the configuration written `nf10×nc20` in Chapter 5. Because both backbones emit a 768-dimensional embedding per frame, this produces an identical per-subject tensor regardless of backbone, so every Tier 1-to-Tier 2 comparison is an isolated backbone swap with the data pipeline, the partitioning, and the per-subject input shape held constant. The scaling investigation of Section 5.7 raises the budget — up to thirty frames per clip, 600 frames per subject — on exactly the same pipeline.

Two complementary subject-disjoint partitioning schemes are used. A fixed three-way train/validation/test split supports multi-seed corroboration runs, and a five-fold subject-disjoint $k$-fold partition of the full cohort replicates the headline numbers. Both are computed at the subject level rather than the clip or frame level, so no subject contributes clips to more than one partition. Subject-level disjointness is the only defensible unit here: the eventual operating point is a referral decision per pregnancy, and clip-level partitioning would leak fetal identity through the strong within-subject correlations between clips of the same heart. There is no path by which a frame of a test subject can have informed training.

The fixed three-way split is supplied with the iFIND release, sized at 2,888, 620, and 620 subjects. The union is exactly 4,128 with zero overlap, and per-class prevalences track the full cohort to within ±0.5 percentage points on every condition column. The locked test partition follows the iFIND-team convention `subject_id` mod $10 \in \{8, 9\}$, inherited so that any external-comparison exercise reads against the same held-out subjects as the partner project [13].

The five-fold partition is generated by $\texttt{fold}(s) = \texttt{subject\_id} \bmod 5$, label-independently of any condition column. It is therefore deterministic, needs no random seed, and is invariant under the cleanup of Section 3.4: a subject lands in the same fold whether or not the CoA gate is applied. Within each fold, training uses the four non-held folds and the held fold serves as the prediction set on which AUROC, AUPR, sensitivity at 95% specificity, PPV at 60% sensitivity, ECE, Brier, and decision-curve net benefit are computed; the supplied validation split serves as the validation set for both protocols. The five per-fold values are aggregated as a mean ± standard deviation across folds, and this per-fold five-fold mean is the canonical discrimination estimand for the dissertation. Per-fold net-benefit values are reported separately as both a five-fold mean and a folds-positive-of-five count (Sections 5.4 and 6.3), because a screening recommendation that

holds only on average across folds is not the same as one that holds on each.

## 3.7 Doppler is not a confounder

A natural worry about acquisition artefacts deserves an explicit test. Sonographers tend to apply Doppler colour-flow imaging when they are investigating a suspected anomaly, so a model that learned Doppler-presence rather than cardiac anatomy could appear to discriminate disease while really detecting a protocol signature. Around 23% of clips in the released cohort carry Doppler colour overlays, at $\leq 4\%$ pixel coverage.

To test the hypothesis we computed, per test-partition subject, the fraction of clips containing a Doppler overlay. Healthy subjects ($n = 524$) average 0.2306; any-unhealthy subjects ($n = 96$) average 0.2331, a difference of $+0.0025$ — twenty times smaller than the 0.05 short-circuit threshold pre-registered as the minimum magnitude below which the confounder direction is treated as absent. Per-condition cells all sit within $\pm 4.6$ percentage points of the healthy baseline, and the auditable-four positive subgroup averages 0.2377 (a deviation of $+0.007$). Doppler clips are therefore retained unfiltered in every reported experiment. The wider threats-to-validity treatment is taken up in Chapter 7.

## 3.8 Label-quality caveats

Three properties of the iFIND release bound the strength of the inferential claims made in Chapters 5–6, and they are stated here rather than buried in a discussion section because they are the genuine ceiling on the work.

**Antenatal suspicion is not postnatal confirmation.** The nine-column labels are antenatal-suspicion labels. With the sole exception of the CoA slice cleaned in Section 3.4, the eight remaining columns retain their antenatal provenance. The clinical collaborator confirmed that antenatal diagnosis approximates ground truth for these eight, but no postnatal-confirmation flag is encoded for them and the residual divergence is unmeasured. We do not extend the CoA discipline to the other eight columns without clinical guidance to do so.

**The healthy class is a not-labelled class, not a verified-negative class.** A zero on every condition column means the subject does not carry a positive antenatal label, not that the subject was examined and ruled out. Free-text annotations reveal a handful of healthy-labelled subjects with notes of non-canonical findings, such as a ventricular septal defect or an echogenic left ventricle. With no postnatal provenance for the negatives, some label noise on the healthy side is unavoidable, and the analysis assumes that this bias is small and roughly symmetric across the discrimination comparisons of Chapter 5. It is a real ceiling, and the dual finding is read with it in mind.

**Small-$n$ per-condition cells are descriptive only.** The two smallest cells in Table 3.3 — aortic stenosis at $n_{\text{pos}} \approx 5$ and pulmonary stenosis at $n_{\text{pos}} \approx 2$ — support neither narrow bootstrap intervals nor sharp between-architecture claims, and Chapter 5 flags their per-condition AUROC as exploratory. Even at $n_{\text{pos}} \geq 6$, per-condition FetalCLIP-versus-DINOv2 deltas are not statistically separable at iFIND scale: every per-condition bootstrap interval pair overlaps. The dual finding of Chapter 6 therefore load-bears on the lumped auditable-four and all-nine tasks, with the per-condition decomposition serving as context rather than as a source of primary clinical claims.

# Chapter 4

# Tier 1: The DINOv2 Baseline

## 4.1 What this chapter has to do

Before a backbone swap can mean anything, there has to be something to swap it against. This chapter builds that reference point. It takes a frozen general-purpose vision backbone, DINOv2, a Vision Transformer trained by self-distillation with no labels [6], attaches five lightweight classifier heads to its embeddings, and evaluates the lot on the 4,128-subject iFIND cohort under the subject-disjoint five-fold protocol that every later chapter reuses without alteration. The result is a fixed bar: an all-nine discrimination ceiling, a clinically-grounded auditable-four figure, a per-condition profile, and a calibration column.

That last column is where the chapter earns its place in the argument. A baseline could be a single number, dutifully recorded and quickly forgotten. This one is not, because the five heads disagree with themselves in a way that turns out to govern the entire dissertation. The head that ranks cases best is not the head that reports the best-calibrated probabilities, and the gap between those two rankings is wide and systematic. I will call this the *rank inversion*, and Section 4.8 measures it with a Brier–Murphy decomposition. The rank inversion is the local substrate of the dual finding that Chapter 6 formalises; everything in Chapters 5 and 6 is read against the comparand fixed here.

## 4.2 A frozen backbone, and why it stays frozen

DINOv2 is an image encoder pretrained on a curated 142-million-image corpus by self-distillation, combining a teacher–student objective with masked-image modelling and a heavy data-curation pipeline [6]. It learns general visual structure without ever seeing a label, and the features it produces are meant to transfer broadly. They carry no knowledge of fetal ultrasound in particular, and nothing about congenital heart disease. The published checkpoint used throughout this tier is `facebook/dinov2-base`, a ViT-B/14 encoder that maps a $224 \times 224$ input frame to a 768-dimensional embedding. That dimension is not incidental: it is the exact interface width that the FetalCLIP backbone also presents in Chapter 5, which is what makes the later swap a clean exchange rather than a confounded change of input shape.

The backbone is used frozen. It is treated as a fixed feature extractor and is never fine-tuned on iFIND. Three considerations drove that decision. The headline question of the dissertation is whether a domain-specialist contrastive backbone extracts more discriminative features than a general-purpose one *on the same head*, and a frozen-versus-frozen comparison is the only way to isolate the backbone as the single moving part. Fine-tuning a ViT-B on a training partition of fewer than three thousand subjects also invites catastrophic over-fitting across the long gap between the pretraining task and ours, a risk Chapter 7 returns to. And the compute envelope for an end-to-end ViT-B fine-tune, with per-subject batches of roughly two hundred frames each carried at 768 dimensions, exceeded what was available for the bulk of the project. Freezing

has a cost: the discrimination ceiling reported here is a frozen-feature ceiling, not the best this architecture could ever reach. Chapter 7 states that limitation plainly.

## 4.3 From clips to a per-subject tensor

The iFIND cohort (Chapter 3) is fed through Tier 1 and Tier 2 by an identical pipeline. Each subject contributes up to twenty clips of ten frames apiece: the canonical `nf10×nc20` budget, ten frames per clip across twenty clips, for two hundred tokens per subject. Every frame passes through the frozen ViT-B/14 tower at $224 \times 224$, and the per-subject output is a $(20, 10, 768)$ embedding tensor — twenty clips, ten frames, 768 features. This is the same shape, and the same width, that the FetalCLIP tower produces in Chapter 5, so the classifier head sees a byte-compatible interface in both tiers. Figure 4.1 sketches the matched-768 swap: both towers feed identically-shaped inputs to the same five heads.

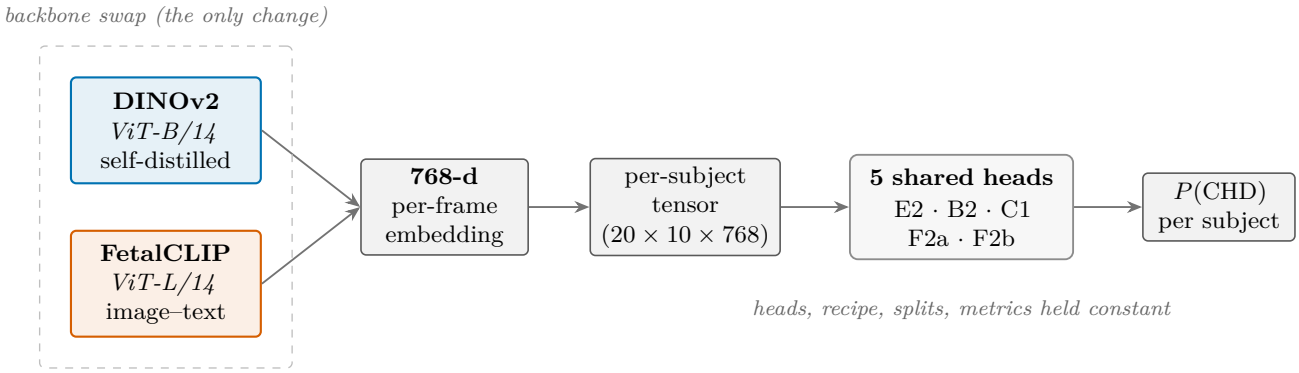


Figure 4.1: The isolated backbone swap. Tier 1 (DINOv2, a self-distilled *ViT-B/14*) and Tier 2 (FetalCLIP, an image–text contrastive *ViT-L/14*) both emit a 768-dimensional per-frame embedding, so each subject is represented by an identically-shaped $(20 \times 10 \times 768)$ tensor and the same five subject-aggregation heads, training recipe, subject-disjoint splits, and evaluation grid apply unchanged to both. Only the frozen backbone differs: every Tier 1 versus Tier 2 contrast in this dissertation is therefore a controlled exchange of representation, not a confounded change of input.

The matched 768-dimensional interface is what licenses the later comparison. DINOv2 reaches 768 natively; FetalCLIP is a wider ViT-L/14 encoder whose internal 1024-dimensional features are projected to 768 before they leave the tower. Holding head, recipe, splits and metrics fixed therefore isolates a bundle of three things that move together: the pretraining *objective* (self-distillation versus image–text contrastive), the pretraining *domain* (general imagery versus fetal ultrasound), and the encoder *capacity* (base versus large). Capacity is the one confound the matched interface does not dissolve, because it lives upstream of the projection. It is named here and resolved later: a capacity-matched DINOv2-Large control arm, discussed in Chapter 7, holds objective and domain fixed while varying only size, which is the clean way to ask whether the backbone effect is really a capacity effect in disguise.

The two-hundred-token budget is a methodological choice, not a hardware limit. Chapter 5 scales the FetalCLIP path up to three times that allowance and finds the discrimination ceiling barely loosens; the Tier 1 path is reported only at the canonical budget, because with FetalCLIP in hand the scaling question was cheaper to ask on the domain-specialist side.

## 4.4 Five heads on one recipe

Five subject-aggregation heads sit on top of the frozen embeddings, ordered here by roughly increasing capacity:

- **E2, the mean-pool linear probe.** The $(20, 10, 768)$ tensor is averaged across both the clip and frame axes into a single 768-dimensional subject vector, and a linear classifier maps it to an unhealthy probability. No attention, no per-clip processing; the parameter count is the embedding width plus a bias, about 1,025.
- **B2, cross-attention pooling.** A single learnable query attends over the two hundred per-frame tokens; the attended vector goes through a linear classifier. About 30K parameters.
- **C1, temporal convolution.** A 1D convolutional stack runs along the ten-frame clip axis, the per-clip representations are pooled across clips, and a linear classifier reads the result. About 70K parameters.
- **F2a, transformer at hidden 64.** A single transformer encoder layer (hidden 64, four attention heads, GELU feed-forward) processes the ten frames of each clip; per-clip predictions are averaged into the subject probability. About 155K parameters.
- **F2b, transformer at hidden 128.** F2a with the hidden width doubled. About 250K parameters.

The training recipe is identical for all five. The optimiser is AdamW with learning rate $5 \times 10^{-4}$ and weight decay $1 \times 10^{-4}$; dropout of 0.3 is applied to the head's hidden representation, with E2 the exception that has none; the batch is 64 subjects; training runs to a maximum of 80 epochs with early stopping on held-out validation negative log-likelihood at a patience of ten. The objective is binary cross-entropy against the all-nine lumped label — positive if any of the nine condition columns is positive (Chapter 3). Folds are the subject-disjoint five-fold mod-5 splits of Chapter 3: for each fold $k$, the test set is the subjects with `subject_id` mod $5 = k$, the validation set is a stratified tenth of what remains, and the rest is training. The five heads are trained and scored on the five folds independently, with no fold crossing into another and no head seeing test subjects during fitting or model selection.

Recipe variations were swept as part of the Chapter 5 scaling work and moved nothing on either backbone, so the canonical recipe above is the only one reported here.

## 4.5 All-nine discrimination: a tight cluster, with a tell

The all-nine task asks the blunt question a screening service starts with — any congenital heart disease versus a healthy heart — using the OR of the nine condition columns as the label. On the full cohort of 4,128 subjects (640 unhealthy, a research-cohort prevalence of 15.50%), the five-fold subject-disjoint AUROCs are reported in Table 4.1.

The discrimination ceiling of the frozen DINOv2 baseline is about 0.80 AUROC, squarely in the 'fair' band of Halligan and colleagues, whose conventional cut-points place $\geq 0.85$ as good, 0.70–0.85 as fair, and $< 0.70$ as poor [2]. The heads are, for practical purposes, indistinguishable on this metric: the gap from B2 to E2 is about 0.015 AUROC, smaller than E2's own fold-to-fold standard deviation of 0.023. Choosing a Tier 1 head on AUROC alone would be choosing on noise. That is the tell, and it is why head selection in this dissertation is routed through the calibration and operating-point evidence of Section 4.8 rather than through this table. All per-fold aggregates in this chapter use the canonical subject-disjoint five-fold estimator.

For orientation, the matching FetalCLIP heads in Chapter 5 sit roughly a tenth of a point higher across the same five splits, with paired-bootstrap intervals on each backbone swap clearing zero. The $\approx 0.80$ ceiling fixed here is the number against which that lift is measured.

| head | all-nine AUROC (5-fold mean ± SD) |
|---|---|
| B2 (cross-attention) | $\mathbf{0.8127 \pm 0.0081}$ |
| F2a (transformer, h64) | $0.8084 \pm 0.0127$ |
| E2 (mean-pool linear probe) | $0.7975 \pm 0.0232$ |

Table 4.1: Tier 1 all-nine binary AUROC for the three heads carried through full five-fold retraining, as per-fold subject-disjoint mean ± standard deviation. The heads sit within a 0.015-AUROC band: B2 holds the point lead, F2a is a whisker behind, and E2, the linear probe, trails the pair by about one fold-standard-deviation. The C1 temporal-convolution and F2b transformer heads, evaluated at the canonical single split rather than retrained across all five folds, fall in the same band and are not tabulated separately here.

## 4.6 Per-condition profile under slicing-i

Lumping nine conditions into one label is convenient for training but hides where the discrimination actually comes from. The per-condition decomposition restratifies the trained binary head's per-subject probability against each condition column under the `slicing-i` convention of Chapter 3: for a condition X, the positives are subjects positive on X (co-morbidity allowed) and the negatives are healthy subjects only. Table 4.2 reports five-fold mean AUROC for E2 — the head this dissertation selects, for reasons that arrive in Section 4.8 — with 95% bias-corrected-and-accelerated bootstrap intervals [23] ($B = 1000$ subject-level resamples, mean of the per-fold intervals) and sensitivity at fixed 95% specificity. The numbers are on the postnatal-confirmation-cleaned cohort of 4,086 subjects (Chapter 3).

| condition | $n_{\text{pos}}$ (fold mean / total) | AUROC (5-fold mean ± SD) | BCa 95% CI | sens@spec= 0.95 |
|---|---|---|---|---|
| HLHS | 16.6 / 83 | $0.870 \pm 0.028$ | $[0.777, 0.924]$ | 0.350 |
| AVSD | 17.2 / 86 | $0.837 \pm 0.057$ | $[0.715, 0.906]$ | 0.372 |
| Pulmonary atresia | 14.8 / 74 | $0.807 \pm 0.029$ | $[0.628, 0.906]$ | 0.379 |
| Tetralogy of Fallot | 23.0 / 115 | $0.803 \pm 0.017$ | $[0.698, 0.878]$ | 0.337 |
| TGA | 25.6 / 128 | $0.795 \pm 0.062$ | $[0.696, 0.867]$ | 0.340 |
| CoA (cleaned) | 11.8 / 59 | $0.791 \pm 0.050$ | $[0.617, 0.895]$ | 0.306 |
| RAA | 21.8 / 109 | $0.760 \pm 0.061$ | $[0.642, 0.852]$ | 0.330 |
| Pulmonary stenosis | 5.6 / 28 | $0.801 \pm 0.054$ | $[0.551, 0.941]$ | 0.244 |
| Aortic stenosis | 5.4 / 27 | $0.799 \pm 0.104$ | $[0.588, 0.950]$ | 0.453 |

Table 4.2: Tier 1 per-condition AUROC for E2 (the DINOv2 mean-pool linear probe) under slicing-i, five-fold subject-disjoint, on the cleaned cohort (4,086 subjects). The lower two rows — aortic and pulmonary stenosis — have a fold-mean of fewer than six positives and are reported descriptively only. Intervals are 95% BCa bootstrap [23], $B = 1000$, taken as the mean of the per-fold intervals.

How to read the table: every interval is bounded away from the 0.5 no-skill line, including the two exploratory stenosis rows whose lower bounds sit at 0.551 and 0.588. The frozen linear probe is doing genuine per-condition work, not merely lumped-binary work that happens to average out. The ordering also has the face validity a fetal cardiologist would expect. Hypoplastic left heart syndrome sits at the top: a structural lesion that distorts the four-chamber view so grossly it is almost unmissable once seen. Right aortic arch sits at the bottom, a subtle anomaly read mainly on the three-vessel-trachea view that ten-frame uniform sampling may not reliably capture, a point Chapter 7 develops.

These per-condition deltas should not be over-read at this many positives. The inferential claim the dissertation defends is the cohort-mean at the lumped level, where the sample-size argument applies to the mean rather than to any single cell. Chapter 5 reports the FetalCLIP per-condition table alongside this one: FetalCLIP wins all nine conditions with ΔAUROC

between +0.07 and +0.12, but the paired BCa intervals overlap, so the per-condition comparison load-bears only at the lumped (all-nine and auditable-four) level. That is the canonical reading throughout.

## 4.7 The auditable-four task

The auditable-four task is the clinically-grounded headline of the dissertation. Its positives are subjects positive on any of TGA, AVSD, Tetralogy of Fallot, or HLHS, regardless of co-morbidity; its negatives are healthy subjects only. These four are the cardiac targets the UK Fetal Anomaly Screening Programme explicitly audits a service against [11], which is precisely why they are the right subgroup to put a screening number on. The auditable-four prevalence in this cohort is about 0.10.

On the canonical pre-cleanup cohort, the five-fold auditable-four AUROC for the headline head E2 is $0.818 \pm 0.022$. Cleaning the cohort by the postnatal-confirmation gating of Chapter 3 lifts the same configuration marginally, to 0.821. The remainder of the dissertation quotes the uncleaned 0.818 as the canonical Tier 1-versus-Tier 2 discrimination figure, to keep it consistent with the uncleaned-cohort net-benefit numbers of Chapter 6; the cleaned variant is consulted once more there, alongside the cleaned temperature-scaling analysis.

The auditable-four AUROC sits about 0.02 above the all-nine ceiling, a small but consistent gain from concentrating on a clinically coherent subgroup and shedding some of the noisier per-condition cells; Chapter 5 sees the same direction on FetalCLIP. The lumped clinical framing wins discrimination on both backbones.

The other reading is the one that sets up the whole argument. Tier 1 auditable-four discrimination, E2 on DINOv2 at $0.818 \pm 0.022$, sits materially below the matching FetalCLIP head, by a margin that clears several fold-standard-deviations and survives paired-bootstrap testing — quantified at its first appearance in Chapter 5. That gap is the evidence for the first clause of the dual finding, that the backbone swap buys discrimination. The companion clause, that the simplest Tier 1 head nonetheless wins the clinical decision, rests on the decision-curve net benefit at the screening threshold $p_t = 0.20$ on this same task, where E2 clears the treat-none reference on all five folds and the higher-AUROC FetalCLIP head sits below it on every fold. Chapter 6 reports that analysis in full; this chapter records only the discrimination half and the calibration mechanism that explains the rest.

Figure 4.2 previews that tension on the all-nine task. How to read it: each panel is one of the five subject-disjoint folds, and each curve is a model's net benefit as the threshold probability runs from 0.02 to 0.40. The screening band is the shaded region from $p_t = 0.10$ to 0.20, and the vertical line marks the headline threshold of 0.20. The signature of the dual finding is that the E2 linear probe sits above the FetalCLIP transformer head across that band on every single fold, even though the transformer head is the better discriminator.

## 4.8 The rank inversion: discrimination is not calibration

The chapter closes on the calibration column the rest of the dissertation leans on. The Brier score, $\mathrm{B} = (1/n)\sum_i (p_i - y_i)^2$, is the mean squared error between a predicted probability and the binary outcome. Murphy's decomposition splits it into three parts [22]: reliability (REL, the squared gap between binned mean prediction and binned empirical frequency — lower is better-calibrated), resolution (RES, the squared gap between binned frequency and base rate — higher is more-discriminating), and uncertainty ($\mathrm{UNC} = \pi(1-\pi)$, a property of the dataset's prevalence $\pi$, not of any model). They combine as $\mathrm{B} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC} + \epsilon$, with $\epsilon$ a small within-bin dispersion residual. Reliability is the term a screening decision cares about: it measures whether a reported probability of 0.2 really corresponds to a one-in-five outcome.

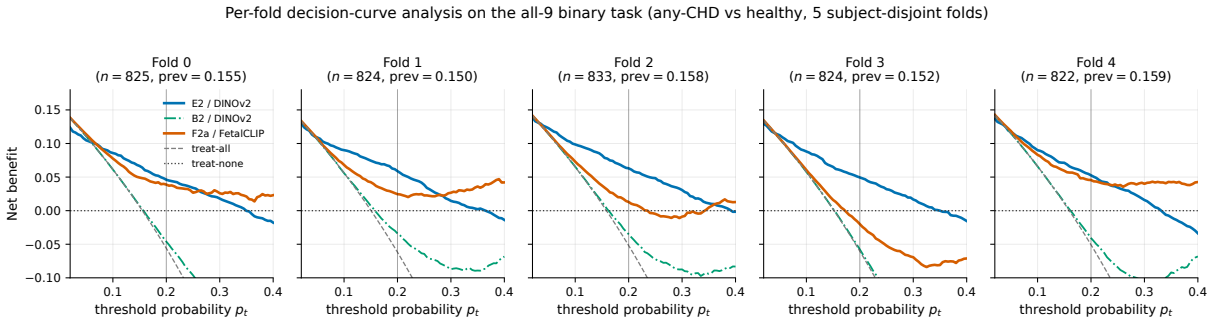


Figure 4.2: Per-fold decision-curve analysis on the all-nine task (any congenital heart disease versus healthy, five subject-disjoint folds). Curves show model net benefit at threshold probability $p_t \in [0.02, 0.40]$. The E2 DINOv2 linear probe (blue) sits above the F2a FetalCLIP transformer head (vermilion) across the screening band $p_t \in [0.10, 0.20]$ on every fold; the B2 DINOv2 cross-attention head (green, dot-dashed) trails both. The treat-all (grey dashed) and treat-none (NB = 0, dotted) reference strategies are drawn for context, and the vertical line marks the headline threshold $p_t = 0.20$. The full clinical-utility treatment is in Chapter 6.

| head | Brier (5-fold mean) | ECE-10 (5-fold mean) | rank |
|---|---|---|---|
| **E2 (mean-pool linear probe)** | $\mathbf{0.1543 \pm 0.0093}$ | $\mathbf{0.1414 \pm 0.0086}$ | **1 (best)** |
| B2 (cross-attention) | $0.1823 \pm 0.0251$ | $0.2679 \pm 0.0458$ | 2 |
| F2a (transformer, h64) | $0.1995 \pm 0.0081$ | $0.2994 \pm 0.0189$ | 3 |

Table 4.3: Tier 1 all-nine Brier score and adaptive expected calibration error (ECE-10) across the five subject-disjoint folds, for the three heads carried through full five-fold retraining. E2 has the lowest Brier and the lowest ECE-10, the latter by roughly a factor of two over the next-best head. The single-split surface preserves this ordering for the two heads not retrained across all folds (F2b and C1 both trail E2 on Brier and ECE).

Table 4.3 reports the result, and it is stark. E2 has the lowest Brier (0.1543) and the lowest ECE-10 (0.1414) of the heads tested. Its ECE-10 is roughly half that of B2 (0.2679) and F2a (0.2994): the transformer head whose AUROC led Table 4.1 is the worst-calibrated head in this table. The operating-point view tells the same story — at fixed 95% specificity on the all-nine task, E2 reaches sensitivity 0.3326, ahead of B2 at 0.3050 and F2a at 0.2915 — so the linear probe does not merely report honest probabilities, it also catches more true positives at the specificity a low-prevalence screen demands.

Set the two tables side by side and the inversion is unmistakable. The discrimination ranking runs B2 $\approx$ F2a $>$ E2; the calibration ranking runs E2 $\gg$ B2, F2a. They are very nearly mirror images. The architecture that wins at ranking cases is the architecture that loses at reporting their probabilities, and it loses at exactly the low-prevalence operating point a screening decision is made at. This is the rank inversion, and it is the local fact the dual finding is built on.

Brier–Murphy says *why* E2 wins, not merely that it does. Its reliability term is the smallest of the three heads, and reliability is the threshold-relevant component of the Brier score. E2's calibration advantage is therefore a genuine reduction in the gap between stated and realised risk, not an artefact of the resolution or uncertainty terms. The mechanism is structural: the linear probe lacks the capacity to inflate mid-range scores, the mechanism developed in full in Section 6.5.

Chapter 5 will show that the FetalCLIP swap flips the discrimination ranking, putting the transformer heads ahead and E2 behind. Chapter 6 will show that the calibration ranking does not flip: the linear probe remains the best-calibrated head and the only configuration whose decision-curve net benefit clears the treat-none reference on all five folds, on either backbone. The

mechanism stated here — head capacity, not the backbone, is the calibration lever — is therefore not a quirk of DINOv2 but a property of the architectural class, which is the point of fixing it in this baseline chapter and the reason its full Brier–Murphy treatment on the auditable-four task lives in Chapter 6.

## 4.9 Summary

The Tier 1 baseline is now fixed. A frozen DINOv2 ViT-B/14 backbone [6] emits 768-dimensional per-frame embeddings into a $(20, 10, 768)$ per-subject tensor, and five subject-aggregation heads — the E2 linear probe, the B2 cross-attention head, the C1 temporal convolution, and the F2a and F2b transformers — are trained on one shared recipe across the subject-disjoint five folds of the 4,128-subject iFIND cohort. All-nine AUROC clusters near 0.80, within about 0.015 across heads; auditable-four AUROC sits at 0.818; per-condition AUROC under slicing-i runs from 0.760 for right aortic arch to 0.870 for HLHS, with every interval clear of the no-skill line.

The load-bearing result is the rank inversion. The head that wins discrimination (B2, with F2a alongside) is roughly the head that loses calibration, and the linear probe that trails on AUROC wins Brier, ECE, and sensitivity at 95% specificity — the calibration win running to about a factor of two on ECE. Chapter 5 swaps the backbone to FetalCLIP and reports the discrimination lift that follows; Chapter 6 turns the calibration column established here into the clinical-utility half of the dual finding.

# Chapter 5

# Tier 2: The FetalCLIP Backbone Swap

## 5.1 What changes when only the backbone changes

Chapter 4 fixed a reference point: a frozen general-purpose backbone, five lightweight heads on one recipe, an all-nine discrimination ceiling near 0.80 AUROC, and a calibration column in which the simplest head — the mean-pool linear probe — reported the most honest probabilities at the screening operating point. This chapter pulls the single lever the dissertation is built around. It replaces DINOv2 with FetalCLIP, a vision–language backbone pretrained on fetal ultrasound, and holds everything downstream of the embedding fixed: the same five heads, the same training recipe, the same subject-disjoint folds, the same metric grid. Whatever moves is attributable to the backbone and the things that travel with it.

The discrimination half of the dual finding is settled here. The swap lifts all-nine AUROC from roughly 0.78 to 0.889 across five subject-disjoint folds, and the clinically-grounded auditable-four AUROC from 0.818 to 0.911. It also reorders the heads: the architecture ranking that Chapter 4 established inverts under the swap, a rank flip set out in Section 5.3. That inversion is the structural hinge between the two halves of the dual finding, and Chapter 6 reads its clinical consequence.

The chapter closes on Q4, the scaling question. Rather than an open-ended search for a positive, it reports a bounded, pre-registered experiment: how discrimination responds as the per-subject token budget grows, whether any of that growth reaches the clinical decision, and which head-side mechanism is the binding constraint on the FetalCLIP ceiling. The short version is that discrimination edges up only gently under the canonical estimator while the clinical net-benefit ceiling does not move at all, and four candidate head-side explanations are falsified in turn. The order-of-magnitude-larger budget that would test the remaining hypothesis is named as a clean open frontier and carried into Chapter 7.

## 5.2 The FetalCLIP backbone, and what travels with the swap

FetalCLIP [7] is a CLIP-style [17] dual-tower vision–language model pretrained on 210,035 fetal ultrasound image–caption pairs. Its image tower is a ViT-L/14 encoder; its text tower is initialised from a published CLIP checkpoint and fine-tuned alongside the image tower under the standard contrastive image–text matching objective. The captions were templated by GPT-4o over structured clinician annotations (view labels, gestational age, pixel spacing) and carry no diagnosis or pathology descriptor. Most of the pretraining corpus came from routine second-trimester scans that the authors note ‘lacked information on fetal health conditions’ [7]; the only cardiac-disease context is roughly two thousand textbook images, about one per cent of the corpus. No part of FetalCLIP’s pretraining is disease-supervised. That matters for Chapter 7,

which attributes the discrimination win to representation quality rather than to leaked disease knowledge.

The backbone is used frozen, exactly as DINOv2 was in Chapter 4. Each subject is processed through the published FetalCLIP image tower at the canonical `nf10×nc20` budget — ten frames per clip across twenty clips, two hundred tokens per subject — and the per-subject output is a $(20, 10, 768)$ embedding tensor. This is byte-for-byte the same shape and width the DINOv2 tower produced, so the five heads see an identical interface on both backbones. The training recipe, the optimiser, the subject-disjoint five-fold split, the bootstrap inference ($B = 1000$, seed 42), and the metric set are all unchanged from Tier 1. Every Tier 1-versus-Tier 2 comparison in this chapter is therefore an isolated backbone swap at a matched 768-dimensional interface.

The word *matched* needs a caveat, and it is the same one Chapter 4 raised. The interface is matched; the towers are not. DINOv2 is a ViT-B with about 86 million parameters reaching 768 dimensions natively; FetalCLIP is a wider ViT-L with about 300 million parameters whose internal 1024-dimensional features are projected to 768 before they leave the tower. Holding head, recipe, splits and metrics fixed isolates a bundle that moves together: the pretraining objective (self-distillation versus image–text contrastive), the pretraining domain (general imagery versus fetal ultrasound), and the encoder capacity (base versus large). Capacity is a named confound, not a controlled variable, because it lives upstream of the projection. Chapter 7 resolves it with a capacity-matched DINOv2-Large control arm that holds objective and domain fixed while varying only size, which is the clean way to ask whether the backbone effect is a capacity effect wearing a different label.

## 5.3 All-nine discrimination: the swap, and a rank flip

The all-nine task is the blunt screening question — any congenital heart disease versus a healthy heart — with the OR of the nine condition columns as the label. The cohort is the full 4,128 subjects (640 unhealthy, a research-cohort prevalence of 15.50%), partitioned into the subject-disjoint five-fold mod-5 splits of Chapter 3. Table 5.1 reports the five-fold AUROC for each head on both backbones, with the per-fold backbone swap as the paired unit.

| head | Tier 1 AUROC (5-fold) | Tier 2 AUROC (5-fold) | ΔAUROC |
|---|---|---|---|
| F2a (transformer, h64) | $0.808 \pm 0.013$ | $\mathbf{0.8906 \pm 0.0185}$ | +0.083 |
| F2b (transformer, h128) | $0.783 \pm 0.020$ | $0.8891 \pm 0.0150$ | +0.106 |
| B2 (cross-attention) | $0.813 \pm 0.008$ | $0.8861 \pm 0.0093$ | +0.073 |
| C1 (temporal conv) | $0.782 \pm 0.018$ | $0.8751 \pm 0.0100$ | +0.093 |
| E2 (mean-pool linear probe) | $0.7975 \pm 0.0232$ | $0.8550 \pm 0.0189$ | +0.058 |

Table 5.1: All-nine binary AUROC for the five heads on each backbone, as five-fold subject-disjoint mean ± SD, with the swap delta in the final column. The lift is uniform across heads, and the headline configurations clear their fold-to-fold spread by several standard deviations under paired-bootstrap resampling [23]. The E2, B2 and F2a Tier 1 figures are the full five-fold means of Table 4.1; the F2b and C1 Tier 1 figures are the canonical single-split values, since those two heads were not retrained across all folds on DINOv2 (Chapter 4).

The same comparison is drawn as a forest plot in Figure 5.1, with the conventional Halligan interpretation bands overlaid so the lift can be read in clinical context. The plot separates the two backbones into two clusters along the AUROC axis, and the bands mark the cut-points of Halligan and colleagues [2] — fair from 0.70, good from 0.80, with 0.90 as the boundary of excellent discrimination.

The swap is a real and uniform lift: every head gains, and the headline heads move the all-nine task from the upper edge of the 'fair' band into 'good' discrimination near 0.89. The

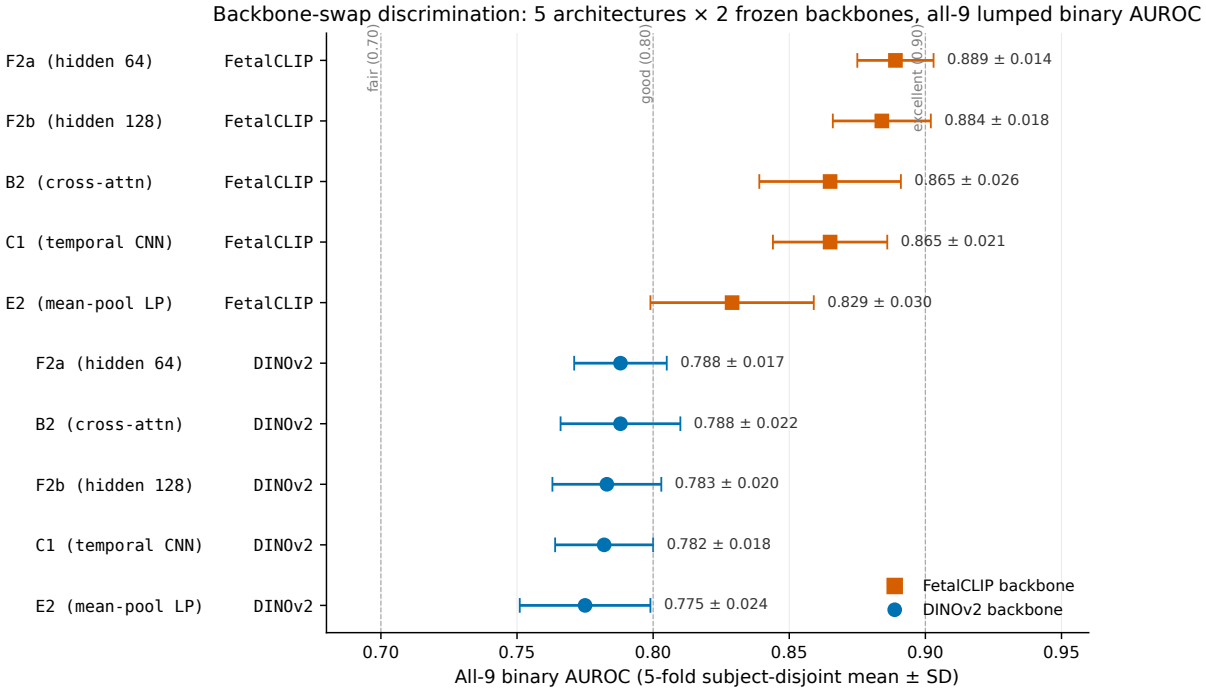


Figure 5.1: Backbone-swap discrimination forest: five subject-aggregation heads on two frozen backbones, all-nine lumped binary AUROC as five-fold subject-disjoint mean ± SD. The FetalCLIP cluster (vermilion) sits at the top of the Halligan 'good' band, approaching 0.89; the DINOv2 cluster (blue) sits a tenth of a point below, in the 'fair'-to-'good' range. Within the FetalCLIP cluster the transformer heads (F2a, F2b) lead and the E2 linear probe trails — the rank flip discussed in Section 5.3. Numbers from Tables 4.1 and 5.1.

lift is also not won by one head alone. F2a, F2b and B2 sit inside a band narrower than their own fold-to-fold spread (0.8906, 0.8891, 0.8861), so the honest reading is the cluster, not the single point. F2a — the one-layer transformer head with a hidden dimension of 64, in the head taxonomy of Chapter 4 — is carried forward as the representative because it is the smallest of the three in parameter count and the most stable across seeds under the multi-seed check reported below. That check, a sweep of five architectures crossed with two embedding normalisations across five seeds on the canonical split, places F2b at $0.891 \pm 0.004$, F2a at $0.889 \pm 0.005$, and the no-normalisation F2a variant at $0.886 \pm 0.007$. The five-fold ± SD and the multi-seed ± SD agree to within 0.005 on the headline configurations, so the 0.889 figure is robust to both data-partition variation and training stochasticity.

The architecture-level result is the rank flip. On DINOv2 the linear probe was the calibration winner and the cross-attention head held the AUROC point lead, with the whole field inside a 0.015 band (Chapter 4). On FetalCLIP the order on AUROC is decisive and reversed: the transformer heads lead at 0.889 and the linear probe trails at 0.855, the worst of the five. The flip is informative in its own right. FetalCLIP embeddings carry discriminative structure that a single linear map over the mean-pooled subject vector cannot extract but a one-layer transformer with mean-of-clip-probabilities readout can. The mechanism behind why this discrimination-superior head nonetheless loses the clinical decision is the business of Chapter 6; the point to carry forward is that the head ordering on the new backbone is the mirror image of the old one.

## 5.4 The auditable-four task

The auditable-four task is the clinically-grounded headline. Its positives are subjects positive on any of TGA, AVSD, Tetralogy of Fallot, or HLHS, regardless of co-morbidity; its negatives are healthy subjects only. These four are the cardiac targets the UK Fetal Anomaly Screening Programme grades a service against [11], which is what makes them the right subgroup for a screening number, and the lumping is deliberate. Per-condition positive counts in the test partition are small enough that single-condition inferences carry wide intervals (Section 5.5), so the auditable-four label both matches how a screening protocol is actually audited and concentrates the statistical power where it can support a claim.

| head | auditable-4 AUROC (5-fold) | DCA NB at $p_t = 0.20$, 5-fold mean | clears treat-none on 5/5 folds? |
|---|---|---|---|
| F2a (FetalCLIP) | $\mathbf{0.911 \pm 0.022}$ | $-0.0372$ | no |
| E2 (DINOv2 linear probe) | $0.818 \pm 0.022$ | $\mathbf{+0.0149}$ | **yes** (5/5) |

Table 5.2: Auditable-four lumped binary task: discrimination (five-fold AUROC) against clinical utility (per-fold decision-curve net benefit at the screening threshold $p_t = 0.20$, uncleaned cohort). FetalCLIP F2a wins discrimination by +0.093; the DINOv2 linear probe E2 wins net benefit, and on every fold. The net-benefit column is the one-row preview of Chapter 6.

The swap lifts auditable-four AUROC from $0.818 \pm 0.022$ on DINOv2 to $0.911 \pm 0.022$ on FetalCLIP, a gain of +0.093 — about four fold-standard-deviations, with non-overlapping paired-bootstrap intervals. The FetalCLIP auditable-four figure also sits about 0.02 above its own all-nine AUROC of 0.889, the same small gain from concentrating on a clinically coherent subgroup that Chapter 4 saw on DINOv2, which we read as the lumping shedding some of the single-condition label noise of Section 5.5. This row is the evidence for the first clause of the dual finding, that the backbone swap buys discrimination.

It is also where the second clause first surfaces, and Table 5.2 sets the two in deliberate tension. The 0.911 discrimination headline does not translate into positive net benefit at the screening threshold: F2a returns −0.0372 and clears the treat-none reference on none of the five folds, while E2 — the DINOv2 mean-pool linear probe of Chapter 4, nearly a tenth of a point worse on AUROC — returns +0.0149 and clears treat-none on all five. The model that wins ranking loses the decision. Chapter 6 dedicates itself to that split, the calibration mechanism behind it, and a temperature-scaling analysis that narrows but does not close it; this chapter records the discrimination half and the single net-benefit row that demands the full treatment.

## 5.5 Per-condition decomposition under slicing-i

Lumping is the right inferential unit, but it hides where the discrimination comes from. The per-condition decomposition restratifies the trained binary head's per-subject probability against each condition column under the `slicing-i` convention of Chapter 3: for a condition X, positives are subjects positive on X with co-morbidity allowed, and negatives are healthy subjects only. Table 5.3 gives the per-condition five-fold AUROC for F2a on FetalCLIP.

Figure 5.2 renders the table as a forest plot ordered by clinical face-validity, with bias-corrected-and-accelerated bootstrap intervals [23]. The vermilion markers are the seven conditions with at least six positives per fold, where the cohort mean is the inferential claim; the grey markers are the two exploratory rows, shown for completeness only.

FetalCLIP wins every one of the nine conditions over the DINOv2 baseline of Chapter 4, with per-condition gains between +0.07 and +0.12. The paired BCa intervals overlap, however, so the per-condition comparison is descriptive and the lift load-bears only at the lumped level, where the all-nine and auditable-four figures of Sections 5.3 and 5.4 carry the claim. The ranking itself has the face validity a fetal cardiologist would expect. HLHS sits at the top: a structural

| condition | $n_{\text{pos}}$ (fold mean) | AUROC (5-fold mean ± SD) |
|---|---|---|
| HLHS | 16.6 | $0.943 \pm 0.034$ |
| Aortic stenosis | 5.4 | $0.915 \pm 0.038$ (exploratory) |
| AVSD | 17.2 | $0.911 \pm 0.030$ |
| Tetralogy of Fallot | 23.0 | $0.903 \pm 0.050$ |
| Pulmonary atresia | 14.8 | $0.899 \pm 0.042$ |
| TGA | 25.6 | $0.898 \pm 0.047$ |
| CoA (cleaned cohort) | 11.8 | $0.879 \pm 0.023$ |
| Pulmonary stenosis | 5.6 | $0.872 \pm 0.067$ (exploratory) |
| RAA | 21.8 | $0.858 \pm 0.043$ |

Table 5.3: Per-condition five-fold AUROC for F2a on FetalCLIP under slicing-i (positives = condition X regardless of co-morbidity; negatives = healthy only). Every condition with $n_{\text{pos}} \geq 6$ has a point estimate above 0.85 and a 95% BCa interval [23] clear of the 0.5 null. The two exploratory rows — aortic and pulmonary stenosis, fewer than six positives per fold — are reported descriptively.

lesion that distorts the four-chamber view so grossly it is almost unmissable once seen. Right aortic arch sits at the bottom, a subtle anomaly read mainly on the three-vessel-trachea view that uniform ten-frame sampling may not reliably capture, which is a point Chapter 7 returns to in the context of the scaling frontier. The CoA row reflects the cleaned cohort, for the reason the next section sets out.

## 5.6 Cohort cleanup: confirmed CoA versus antenatal suspicion

Coarctation of the aorta differs from the other eight conditions in one clinically important way: its antenatal diagnosis has materially lower positive predictive value, with roughly half of subjects labelled CoA-positive from the antenatal scan not confirmed at birth. The canonical labels carry no postnatal-confirmation column. After an empirical investigation of the filename token field, a postnatal-confirmation flag was decoded from the relevant filename token and used to construct a cleaned cohort that removes antenatal-CoA-suspicion-without-postnatal-confirmation subjects from both training and evaluation. The cohort shrinks from 4,128 to 4,086 subjects, and the CoA-positive count over the full cohort drops from 101 to 59.[1]

On the cleaned cohort, F2a's five-fold CoA AUROC is $0.879 \pm 0.023$, up from roughly 0.85 on the uncleaned cohort. The +0.03 gain is attributable entirely to label cleanup, not to any change in the model. The auditable-four AUROC is unchanged at $0.911 \pm 0.022$, since CoA is not among the auditable-four targets. The substantive consequence for this chapter is narrow: the CoA row of Table 5.3 reflects the postnatal-confirmed cohort while the other eight rows reflect the full cohort. Chapter 7 frames the wider label-quality issue — the healthy class includes an unknown fraction of subjects with minor undiagnosed disease — as a limitation and a future-work direction.

## 5.7 The scaling experiment (Q4)

Token budget was the pre-registered scaling axis: if more tokens per subject sharpen the FetalCLIP transformer head, the discrimination curve carries a positive slope and the dissertation measures it; if the curve is shallow, the dissertation reports a bounded ceiling with a mechanistic account of what binds it. The experiment is deliberately scoped. It measures the discrimination

[1]The clinical motivation for the CoA-specific cleanup, and the residual-risk caveat that the remaining eight condition labels are antenatal-suspicion rather than postnatal-verified, are recorded in Chapter 3.

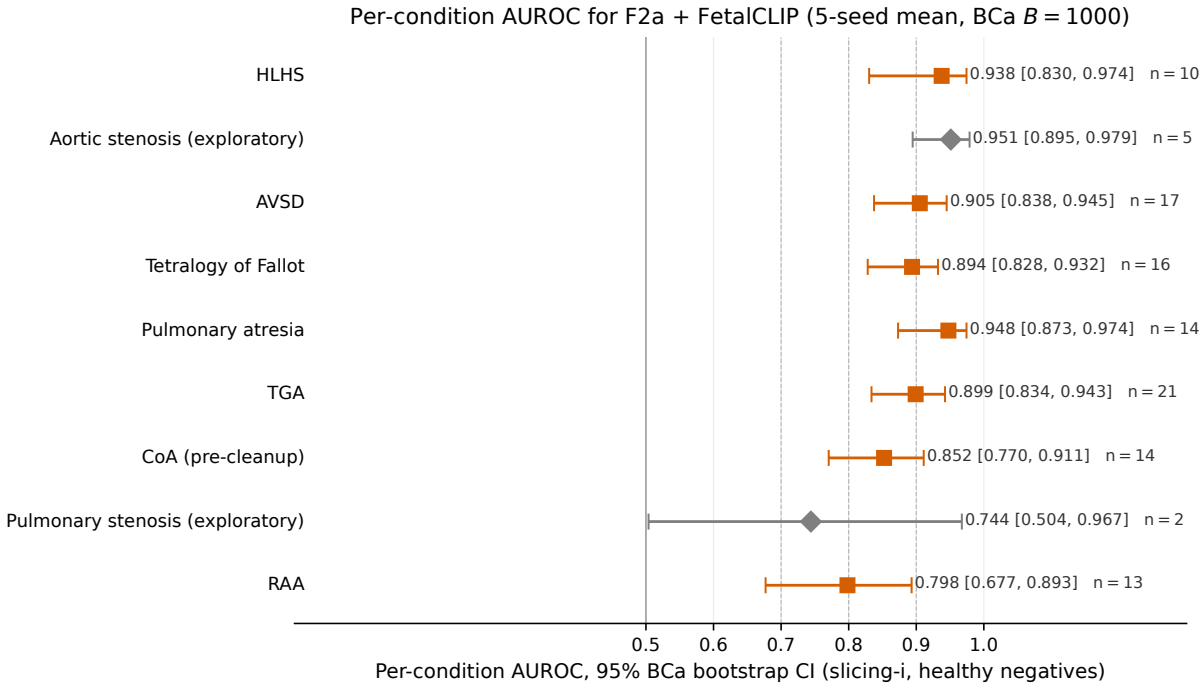


Figure 5.2: Per-condition AUROC forest for F2a on FetalCLIP under slicing-i (positives = condition X regardless of co-morbidity; negatives = healthy only). Bars are 95% BCa bootstrap intervals [23] ($B = 1000$ subject-level resamples per seed, mean of per-seed intervals). Vermilion squares: the seven conditions with $n_{\text{pos}} \geq 6$, for which the cohort mean is the inferential claim. Grey diamonds: the two exploratory rows with $n_{\text{pos}} < 6$, descriptive only. HLHS sits at the top and RAA at the bottom.

gradient under the canonical five-fold estimator, asks whether any of that gradient reaches the clinical decision, and falsifies four candidate head-side explanations for the ceiling in turn. The asymptotic regime that the experiment cannot reach at three times the budget is named at the end as the open frontier.

### 5.7.1 Discrimination under the canonical estimator

The first scaling increment re-extracts the FetalCLIP per-clip embeddings at `nf30×nc20` — thirty frames per clip across twenty clips, six hundred tokens per subject, three times the canonical two-hundred-token allowance. A second increment, `nf10×nc40`, doubles the budget along the clip axis rather than the frame axis. Each is evaluated with the same F2a head, the same recipe, and the same five subject-disjoint folds, so the discrimination gradient is read on the canonical estimator throughout.

The discrimination response is a modest, consistent lift rather than a flat line or a fall. On the canonical five-fold estimator the baseline `nf10×nc20` sits at $0.891 \pm 0.017$ all-nine and $0.911 \pm 0.021$ auditable-four. Doubling the clip budget to `nf10×nc40` raises these to $0.908 \pm 0.004$ and $0.930 \pm 0.010$; tripling the frame budget to `nf30×nc20` raises them to $0.904 \pm 0.009$ and $0.923 \pm 0.006$. Both increments and both tasks move the same way, by roughly $+0.013$ to $+0.019$ AUROC — about one fold standard deviation. There is no step change of the kind a saturating-then-breaking curve would show, but neither is the response flat or degrading: more tokens buy a small, repeatable sharpening of the ranking. The next budget corner, `nf30×nc40` at roughly six times the canonical budget, is in extraction and will be reported on the same per-fold estimator when it lands.

Figure 5.3 plots that discrimination response on a log-spaced token-budget axis with per-fold error bars, alongside the flat net-benefit overlay developed below. The upper traces are the

all-nine and auditable-four AUROC trends with fold-wise error bars; the lower trace is the per-fold net-benefit ceiling, which the next subsection shows does not move at any budget.

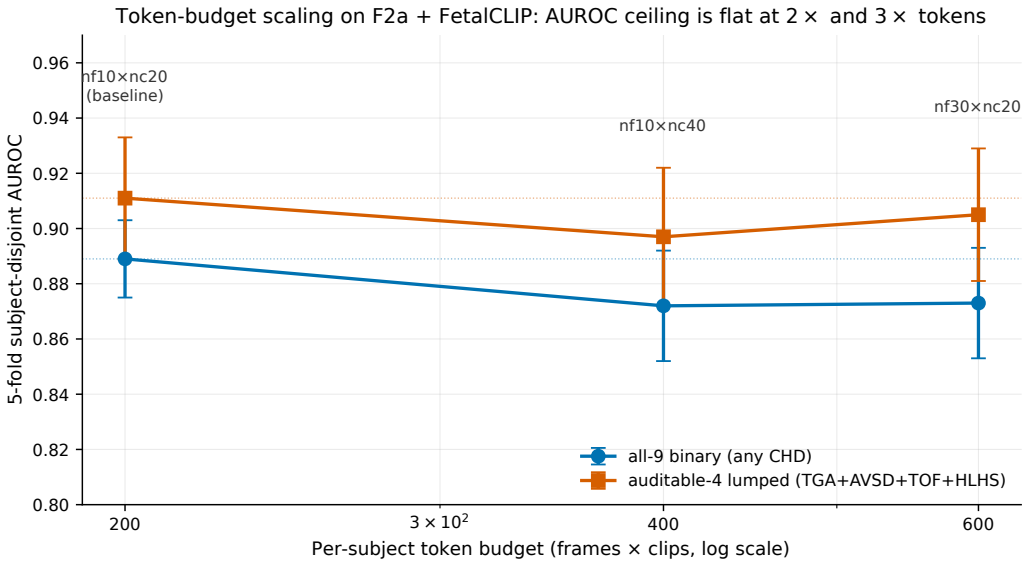


Figure 5.3: Token-budget scaling for F2a on FetalCLIP, on the canonical five-fold estimator with per-fold error bars. The all-nine (blue) and auditable-four (vermilion) AUROC trends rise modestly across the budgets tested (200, 400, 600 tokens per subject), by roughly +0.013 to +0.019 on each increment — about one fold standard deviation. Overlaid is the per-fold decision-curve net benefit at $p_t = 0.20$, which stays negative on every fold (0/5) at every increment, so none of the discrimination lift reaches the screening operating point.

### 5.7.2 The clinical-utility ceiling is invariant to budget

The discrimination gradient is the less important of the two results. The clinical one is sharper and entirely consistent: the decision-curve net benefit of F2a at the screening threshold $p_t = 0.20$ stays negative on every fold, zero of five, at every budget increment tested. Tripling the token budget does not move the FetalCLIP head above the treat-none reference, and neither does doubling it along the clip axis. The screening ceiling is invariant to budget in a way the discrimination gradient is not. Whatever modest ranking improvement more tokens buy, none of it reaches the operating point where the screening decision is made — which is the second clause of the dual finding restated against the scaling axis, and the reason the experiment reports net benefit per fold rather than AUROC alone.

### 5.7.3 A four-way falsification of head-side mechanisms

The invariant ceiling raises a question the chapter has to answer before attributing it to the backbone: is the ceiling a property of the FetalCLIP image encoder, already saturated at the canonical budget, or a property of the F2a head that happens to sit on top of it? Four head-side hypotheses were tested in parallel, each capable of explaining the ceiling if it were a head artefact. Table 5.4 states each as a branch with its verdict and the mechanism behind it.

Three of the four branches are flat nulls, and the two that do anything point the wrong way: a more expressive clip-then-subject aggregation does worse rather than better, starved of gradient because a subject-level loss back-propagates roughly twenty times less signal into the per-clip encoder than a clip-level loss would, and tightening the view-aware filter strictly lowers fold-mean AUROC even though the filter sanity-passes as label-agnostic. Neither tuning the recipe, respecting the per-clip structure, nor removing low-relevance clips loosens the ceiling at

| branch | hypothesis | verdict | mechanism |
|---|---|---|---|
| head capacity | a higher-capacity head absorbs the extra tokens | partial | a four-layer transformer head at 3× budget gives the best aggregate calibration in the study (Brier 0.091, REL 0.034) but does not reach the screening net benefit |
| training recipe | the head is under-tuned | null | a one-axis sweep over learning rate, dropout, weight decay, batch size, epochs and warmup (15 fold-0 trials plus a 3-config five-fold confirmation) finds no recipe that beats the anchor |
| hierarchical aggregation | flat per-clip pooling dilutes the signal | null | more expressive clip-then-subject attention variants do worse; the per-clip encoder is starved of gradient under a subject-level loss, receiving roughly 20× less than a clip-level loss would supply |
| view-aware selection | low-relevance clips add noise | null | hard top-K filtering on text-encoder cosine similarity to cardiac-view prompts is monotonically hurtful: fold-mean AUROC falls as K tightens |

Table 5.4: The four head-side hypotheses tested against the invariant scaling ceiling. Three are null: recipe tuning, hierarchical aggregation, and view-aware clip selection all fail to loosen the ceiling, and two of the three actively hurt. The fourth, head capacity, is partial: added depth buys aggregate calibration but does not transfer to the screening threshold. The ceiling is therefore not a head artefact, which isolates the binding constraint to the image-representation side under a data-starved subject-level signal.

six hundred tokens. Only added head depth moves a number, and it moves the wrong one — so the head side is exhausted, and by elimination the binding constraint sits in the FetalCLIP representation itself, not in anything the head does on top of it.

### 5.7.4 The four-layer head: aggregate calibration without the decision

The head-capacity branch is the one that moves a number, so it earns its own paragraph. F2a_4layer replaces F2a's single transformer encoder layer with four stacked layers, about 255K parameters against F2a's 155K. At three times the budget (`nf30×nc20`) it produces the best aggregate calibration in the entire study — the Brier and reliability figures of Table 5.4, lower than any other configuration on either backbone, including the DINOv2 linear probe. By the aggregate measure the Brier–Murphy decomposition captures, this is the best-calibrated model in the dissertation.

It does not win the decision. At the screening threshold its per-fold net benefit on auditable-four is $-0.0108$, positive on three of five folds — a real improvement on the F2a baseline's $-0.0372$ at zero of five, but still below the treat-none reference and below the DINOv2 linear probe. Added depth opens a genuine aggregate-calibration channel without delivering that calibration to the threshold where the screening decision is made. Chapter 6 unpacks why aggregate calibration and threshold calibration are different objects, using exactly this configuration as the case study; the relevant point here is that the one head-side lever that moves at all moves aggregate calibration, not screening utility, which is why the head-capacity branch is logged as partial rather than as a win.

### 5.7.5 An estimator correction, caught in review

A first-pass writeup of the head-capacity branch reported that scaling had crossed the dual-finding clinical-utility leader: a net benefit at $p_t = 0.20$ of $+0.0415$ for the four-layer head,

compared against the linear probe's +0.0149. A second-pass verification found the comparison was apples-to-oranges. The favourable numerator had been computed by pooling predictions across the five folds onto the locked test split and scoring a single net benefit there, while the reference was the canonical per-fold estimate — net benefit computed within each fold and averaged. The two estimands give materially different numbers on the same predictions, as Table 5.5 shows. The error was caught in internal review and corrected before the claim entered the dissertation.

| estimator | E2 NB at $p_t = 0.20$ | F2a_4layer (nf30) NB at $p_t = 0.20$ |
|---|---|---|
| canonical per-fold, auditable-4 | +0.0149 (5/5 folds) | −0.0108 (3/5 folds) |
| pooled-on-test-split | +0.0540 | +0.0415 |

Table 5.5: The two estimands behind the corrected claim. Under the canonical per-fold method the linear probe wins; the first-pass 'scaling exceeds the baseline' result had compared a pooled-on-test-split estimate against a per-fold one, which manufactured a spurious crossing. Under either consistent estimator the linear probe leads.

Under either consistent estimator the linear probe leads, so the corrected reading is unambiguous: the four-layer head improves on the F2a baseline but does not unseat the dual-finding leader. Per-fold net benefit is the canonical estimand for a $k$-fold setting, because pooling predictions across folds before scoring is unbiased only if the fold-wise models are exchangeable and their predictions are jointly calibrated — and neither holds when each fold is trained on a different subset and calibrated on its own validation split. The pooled surface is reported elsewhere only where it is labelled as such. The same convention governs the scaling-curve discrimination figures of Section 5.7.1: they are quoted as per-fold five-fold mean ± SD so that the scaled-budget AUROC is directly comparable with the per-fold baseline, rather than read off a pooled-bootstrap surface on the locked test split. Chapter 6 adopts the per-fold convention throughout.

### 5.7.6 The open frontier

Three times the canonical budget is enough to bound the response and isolate the binding constraint, and not enough to settle whether the constraint is fundamental. With the head side exhausted by the four-way falsification above, the binding constraint sits in the FetalCLIP representation itself, under a data-starved subject-level training signal — which leaves exactly one hypothesis untested at the scale this chapter could reach: whether an order-of-magnitude-larger budget breaks the ceiling.

The pre-registered frontier is the roughly twenty-times-baseline regime (`nf60×nc60`), carried into Chapter 7 with a falsifiable prediction. If the FetalCLIP representation is scanning-style-bound — limited by what a frozen general fetal-ultrasound encoder captures about subtle cardiac views rather than by how many of those views it is shown — the conditions whose signal lives on views the clips may not sample, right aortic arch and coarctation, should stay near chance even at the larger budget, with mean AUROC at or below 0.4. If instead the constraint is the number of views the head sees, those conditions should climb. The decision rule is symmetric: a lift is a contribution, and a saturation is the bounded answer that more tokens are not the lever, which is the experiment's result rather than a gap in it.

## 5.8 Chapter summary

The FetalCLIP backbone swap lifts the all-nine task ceiling from roughly 0.78 to 0.889 and the auditable-four headline from 0.818 to 0.911, replicated across five subject-disjoint folds and corroborated under two stochastic axes, data partitioning and training initialisation. The

smallest transformer head, F2a, is the headline, with F2b and B2 inside a band narrower than the fold spread; the head ordering inverts across the swap (Section 5.3), a flip Chapter 6 reads for its clinical consequence. Per-condition AUROC across the seven non-exploratory conditions favours FetalCLIP on every condition, load-bearing only at the lumped level, and a postnatal-confirmation cleanup lifts the CoA-specific AUROC by about 0.03 with no change to any model. The scaling experiment bounds Q4: under the canonical per-fold estimator discrimination edges up only gently while the screening net-benefit ceiling stays negative on every fold at every budget, and four head-side mechanisms are falsified — recipe, hierarchical aggregation and view-aware selection are flat nulls, and added head depth buys the best aggregate calibration in the study without reaching the decision. The remaining hypothesis, the roughly twenty-times-baseline regime, is the pre-registered open frontier carried into Chapter 7. The clinical-utility consequence of these discrimination results, the central contribution of the dissertation, is the subject of Chapter 6.

# Chapter 6

# Clinical Utility: The Decision-Curve Bridge

## 6.1 From discrimination to decision

Chapter 5 settled the discrimination question: swapping DINOv2 for FetalCLIP lifts auditable-four AUROC by roughly 0.09, and the smallest transformer head on the domain backbone (F2a) leads that contest. A higher AUROC is a better ordering of subjects by risk. It is not, by itself, a better screening decision. The decision a fetal-cardiac model is actually asked to make has a fixed shape: refer or wave through, at a particular operating point, under a roughly one-per-cent prior. Ranking quality and decision quality can part ways at that point. This chapter formalises the second half of the dual finding by asking the decision question directly, and the answer inverts the discrimination ranking.

What follows is the second half of the dual finding of Chapter 1. This chapter does the work behind three of its clauses: it shows that the DINOv2 linear probe (E2) wins clinical net benefit at the screening threshold, that the head architecture rather than the backbone is the calibration lever that decides that contest, and that a single-parameter temperature rescaling narrows the gap at the mean but does not survive the per-fold replication bar, so it qualifies the mechanism rather than overturning it.

## 6.2 Decision-curve analysis and the screening threshold

The area under the receiver-operating-characteristic curve (AUROC) is the field default for a binary medical classifier, and Chapter 1 already gave the reason it is the wrong instrument for a screening decision: it is invariant to any monotone rescaling of the scores, so two models with identical AUROC can carry arbitrarily different calibration and therefore arbitrarily different value at the one threshold that governs a referral [2, 3]. Decision-curve analysis (DCA) repairs this by scoring a model at a chosen threshold rather than averaging over all of them [3]. For each candidate threshold probability $p_t$ it reports the *net benefit* — true positives identified, minus false positives discounted by the odds of that threshold:

$$\mathrm{NB}(p_t) = \frac{\mathrm{TP}}{n} - \frac{\mathrm{FP}}{n} \cdot \frac{p_t}{1 - p_t}.$$

The discount factor $p_t/(1 - p_t)$ encodes the clinical trade-off the threshold implies. Setting $p_t = 0.20$ asserts that one false-positive referral costs a quarter of one missed case, or equivalently that a clinician would act on any subject whose posterior risk exceeds twenty per cent. A useful model must clear two reference strategies at the chosen threshold: *treat-none* (which has zero net benefit at every $p_t$) and *treat-all* (net benefit = prevalence $-$ (1 $-$ prevalence) $\cdot\, p_t/(1 - p_t)$). A

model whose curve dips below either reference is, at that threshold, worse than deciding without it.

Where to read the curve is fixed by the clinical operating point of fetal echocardiography. Expert four-chamber-plus-outflow screening runs at a specificity near 0.9998 with sensitivity around 0.89 [12]; at the roughly one-per-cent prevalence of congenital heart disease, even a specificity of 0.95 would flag one healthy mid-trimester scan in twenty and overwhelm the referral pathway. This dissertation reports at a pre-registered specificity of 0.95 — a deliberately sub-clinical screening regime, adopted because the cohort lacks the positive count to resolve sensitivity differences at 0.99 with any precision — and evaluates DCA over the grid $p_t \in \{0.10, 0.15, 0.20\}$, with $p_t = 0.20$ as the headline. The looseness of the operating point is a limitation the chapter returns to in Section 6.7; it does not affect the internal comparison between models, which is the chapter's load-bearing claim.

Every net-benefit number in this chapter is a per-fold estimate. Each of the five subject-disjoint folds yields one net-benefit value, and the five-fold mean ± SD is the inferential claim. The methodology audit in Chapter 5 (Section 5.7.5) established why this matters: pooling predictions across folds before computing net benefit is a different and weaker estimand when each fold's model is trained on a different subset and calibrated on a different validation split. Per-fold net benefit is the canonical decision estimand throughout, and pooled surfaces appear only where they are labelled as such.

## 6.3 The per-fold decision curve

The centrepiece is the per-fold net benefit on the auditable-four lumped binary task — positives are TGA, AVSD, Tetralogy of Fallot, or HLHS; negatives are healthy subjects — across the screening grid. Table 6.1 gives the five-fold mean for the headline configurations on the uncleaned cohort, with the per-fold count of folds clearing treat-none in parentheses.

| head | $p_t = 0.10$ | $p_t = 0.15$ | $p_t = 0.20$ | clears treat-none, 5/5 folds? |
|---|---|---|---|---|
| F2a (FetalCLIP) | +0.0170 (5/5) | −0.0160 (1/5) | −0.0372 (0/5) | no |
| F2a_4layer (FetalCLIP) | +0.0286 (5/5) | +0.0036 (3/5) | −0.0108 (3/5) | no |
| **E2 (DINOv2 linear probe)** | **+0.0497** (5/5) | **+0.0321** (5/5) | **+0.0149** (5/5) | **yes** |

Table 6.1: Per-fold decision-curve net benefit on the auditable-four lumped binary task (uncleaned cohort), reported as the five-fold mean with the number of folds clearing the treat-none reference in parentheses. E2, the DINOv2 mean-pool linear probe, clears treat-none on every fold at every threshold in the screening grid; both FetalCLIP heads fall below it at $p_t = 0.20$.

The pattern is unambiguous. At the headline threshold $p_t = 0.20$, the DINOv2 linear probe returns a net benefit of +0.0149 and does so on all five folds, while the higher-AUROC FetalCLIP transformer head returns −0.0372 and clears treat-none on none of them. E2 is the only configuration whose net benefit exceeds the treat-none reference on every fold at the screening threshold, and it leads at $p_t = 0.10$ and $p_t = 0.15$ as well. The second clause of the dual finding is operationalised here: the model that wins discrimination loses the clinical decision. The four-layer transformer head, the secondary discrimination result of Chapter 5, narrows the gap — it lifts the FetalCLIP net benefit from −0.0372 to −0.0108 and moves three of five folds above treat-none — but it stays below the reference, and the finding holds.

Figure 6.1 shows the full curves. Read each panel as one subject-disjoint fold: the horizontal axis is the threshold $p_t$, the vertical axis is net benefit, and the vertical line marks the headline threshold $p_t = 0.20$. The DINOv2 linear probe (blue) sits above the FetalCLIP head (vermilion) across the screening band on every fold, and the FetalCLIP curve drops below the treat-none reference at $p_t = 0.20$ in all five.

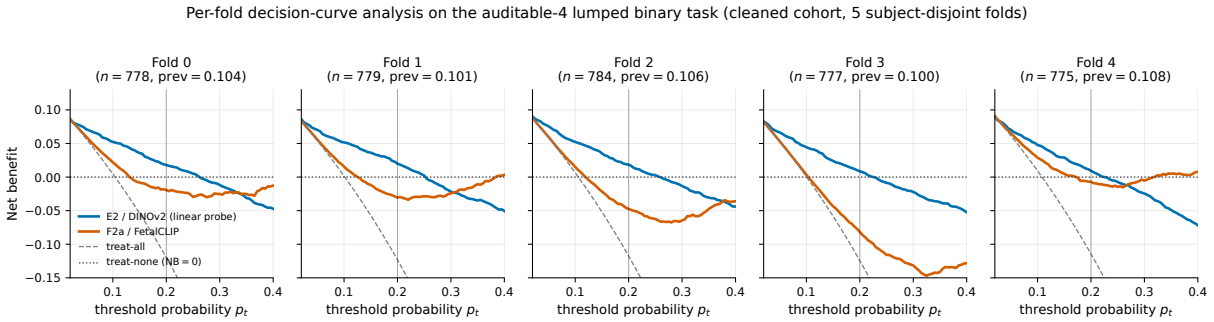


Figure 6.1: Per-fold decision-curve analysis on the auditable-four lumped binary task (uncleaned cohort, five subject-disjoint folds). Each panel reports net benefit against threshold probability $p_t \in [0.02, 0.40]$ for E2 / DINOv2 (blue), F2a / FetalCLIP (vermilion), treat-all (grey dashed), and treat-none (NB = 0, dotted). The vertical grey line marks the headline screening threshold $p_t = 0.20$. On every fold E2 clears treat-none and F2a falls below it at $p_t = 0.20$, the graphical form of the dual finding's second clause.

The all-nine binary task shows the same ordering with larger absolute net benefits and a smaller gap: at $p_t = 0.20$, E2 returns +0.0540 on all five folds and the FetalCLIP head returns +0.0204, positive but roughly 0.034 behind. Restricting to the auditable-four targets *sharpens* the finding rather than weakening it. Auditable-four positives are a low-prevalence subset of the all-nine positives, and at a fixed threshold net benefit is more sensitive to calibration the lower the prevalence, so the FetalCLIP head's miscalibration costs it more on the narrower, more clinically pointed task. Treat-all is not a serious comparator here: at $p_t = 0.20$ on the auditable-four task its net benefit is roughly −0.120, dominated by treat-none, because intervening on every subject at this prevalence is worse than intervening on none. The reference that matters is treat-none, and any configuration sitting below zero at the threshold adds nothing over the default of not screening with the model at all.

## 6.4 Why the linear probe wins: a Brier–Murphy decomposition

The third clause attributes the calibration advantage to the head, not the backbone. The Brier–Murphy decomposition [22] is the natural test, because it factors the Brier score into the parts a clinician cares about: Brier = REL − RES + UNC. Reliability (REL) measures the squared gap between a bin's mean prediction and its observed frequency, so lower is better-calibrated. Resolution (RES) measures how far each bin's frequency departs from the base rate, so higher is more discriminating. Uncertainty (UNC) is fixed by the prevalence alone, $\pi(1-\pi)$, and is a property of the task rather than the model. Table 6.2 gives the decomposition on the auditable-four five-fold predictions.

| head | Brier | REL | RES | UNC (base rate) |
|---|---|---|---|---|
| F2a (FetalCLIP) | $0.110 \pm 0.026$ | $0.050 \pm 0.028$ | $0.031 \pm 0.003$ | 0.093 |
| F2a_4layer (FetalCLIP, nf30×nc20, 3× budget) | $\mathbf{0.091 \pm 0.028}$ | $\mathbf{0.034 \pm 0.030}$ | $0.035 \pm 0.003$ | 0.093 |
| E2 (DINOv2 linear probe) | $0.142 \pm 0.011$ | $0.063 \pm 0.008$ | $0.015 \pm 0.002$ | 0.093 |

Table 6.2: Brier–Murphy decomposition on the auditable-four five-fold predictions. UNC $= \pi(1-\pi)$ with auditable-four prevalence $\pi \approx 0.104$, so it is constant at 0.093. The four-layer head wins the aggregate-calibration metrics (lowest Brier, lowest REL), but this row is the three-times-budget extraction (nf30×nc20), not the baseline allowance, and is labelled accordingly. E2 has the worst REL of the three yet wins net benefit at $p_t = 0.20$ in Table 6.1.

Figure 6.2 renders the same numbers: the left panel is the five-fold mean Brier score, and the right panel splits each head's Brier into its Murphy components. Read the right panel as a stacked bar where REL (vermilion) is the part the model can fix by calibrating, −RES (green) is the part it earns by discriminating, and UNC (sky blue) is the irreducible floor that the prevalence sets.

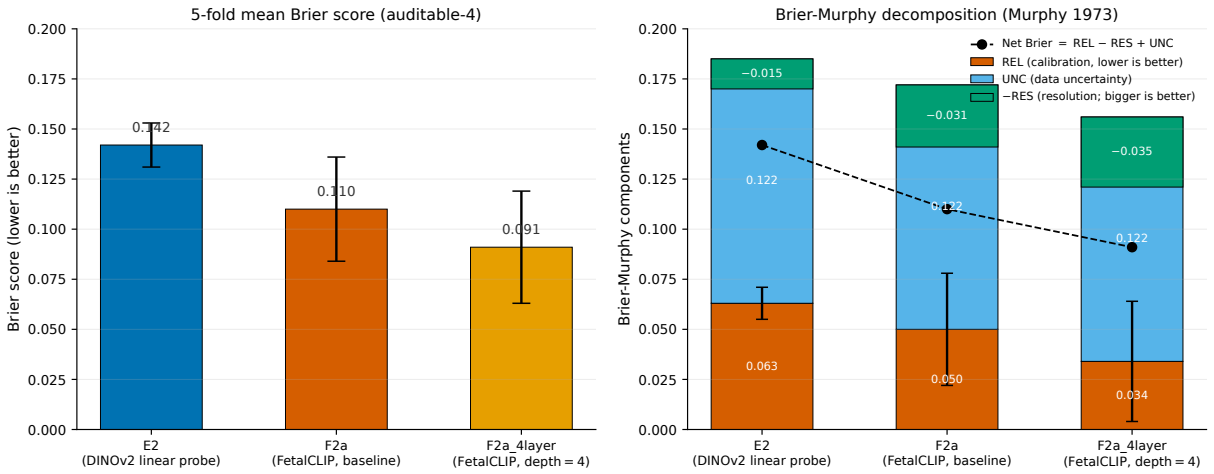


Figure 6.2: Brier–Murphy decomposition on the auditable-four five-fold predictions. **Left:** five-fold mean Brier score (lower is better). **Right:** the Murphy decomposition Brier = REL − RES + UNC. REL (calibration; vermilion) and −RES (resolution; green) are model properties; UNC (sky blue) is the prevalence-determined floor $\pi(1-\pi) \approx 0.093$ at auditable-four prevalence $\pi \approx 0.104$. The four-layer head has the lowest Brier and REL but, as the prose explains, that is aggregate calibration at a three-times budget; E2 has the highest REL and still wins net benefit at the screening threshold (Figure 6.1). Numbers from Table 6.2.

There is an apparent contradiction to resolve. By REL, the best-calibrated head in the entire study is the four-layer FetalCLIP transformer at 0.034, and the baseline FetalCLIP head at 0.050 also beats E2's 0.063. By that measure the linear probe is the *worst*-calibrated of the three. Yet the linear probe wins net benefit at $p_t = 0.20$. The two metrics are not in conflict, because they measure calibration over different regions. The Brier–Murphy reliability term averages over the whole $[0, 1]$ prediction interval — it is *aggregate* calibration — whereas decision-curve net benefit isolates calibration at the single threshold of interest, which is *threshold* calibration. The four-layer head's extra capacity buys better calibration on average, but it spends much of that capacity placing predictions in the $[0.10, 0.30]$ band straddling the decision boundary, where the model is genuinely uncertain and the predictions resolve in neither direction cleanly. The aggregate gain therefore does not reach the threshold where the decision is made.

## 6.5 Where the probability mass sits

The reliability diagrams make the threshold-versus-aggregate distinction concrete by showing where each head puts its probability mass. The canonical mechanism is this: the linear probe maps mean-pooled embeddings through a single weight vector and a sigmoid, so it has no capacity to inflate mid-range scores, and its mass therefore settles below the screening threshold where a low-prevalence decision needs it.

Figure 6.3 shows the consequence. Read the top row as the DINOv2 linear probe and the bottom row as the FetalCLIP head; the left column is a stacked histogram of predicted probabilities, healthy in blue and auditable-four-positive in vermilion, with the dotted line at the auditable-four prevalence; the right column is the reliability curve against the diagonal, with the bin masses as grey bars. The linear probe places almost all of its healthy mass below the

prevalence line, so few healthy subjects ever cross the threshold. The FetalCLIP head spreads its mass across $[0.0, 0.6]$ with a peak straddling $p_t = 0.20$, which is exactly where false positives are manufactured at the screening operating point.

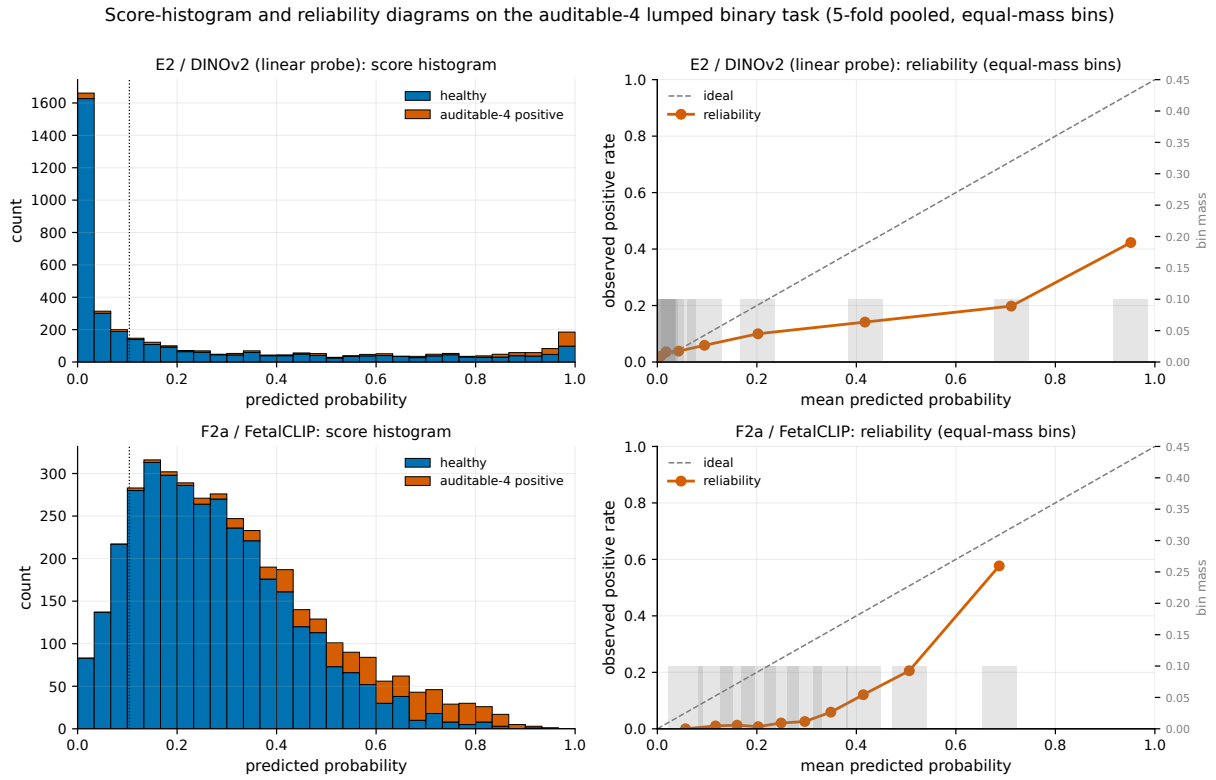


Figure 6.3: Score histograms and reliability diagrams on the auditable-four lumped binary task, five-fold predictions pooled across subject-disjoint folds, equal-mass bins. **Top row:** E2 / DINOv2 linear probe. **Bottom row:** F2a / FetalCLIP. **Left column:** stacked histogram of predicted probabilities for healthy (blue) and auditable-four-positive (vermilion) subjects, with the dotted line at auditable-four prevalence. **Right column:** reliability curve (vermilion) against the identity (grey dashed), with bin masses (grey bars, right axis). E2 concentrates its healthy mass below the prevalence line; F2a spreads its mass across $[0.0, 0.6]$ with a peak straddling the screening threshold $p_t = 0.20$.

This specialises the third clause rather than restating it. The linear probe is the *threshold*-calibration mechanism. Added transformer depth is a genuine *aggregate*-calibration mechanism — it lowers Brier and REL — but it does not deliver that calibration to the screening threshold, and a deployment that selected its head on aggregate Brier alone would pick the head that loses the decision. Chapter 7 carries this distinction into the clinical-deployment discussion.

## 6.6 Temperature scaling: a qualified bridge

If the FetalCLIP head's high AUROC reflects a discrimination capacity that is merely miscalibrated — correct rank order, wrong absolute value at the threshold — then a post-hoc rescaling should recover the lost net benefit. The standard tool is temperature scaling [5]: a single scalar $T$ applied to the logits before the sigmoid. We fit it by minimising negative-log-likelihood on out-of-fold predictions for each held-out fold (concatenating the other four folds, fitting one $T$, applying it unchanged to the held fold), so the evaluation set never participates in selecting its own temperature. The fit is replicated on both the uncleaned cohort and the labelling-cleaned cohort of Section 5.6, with an in-fold half-split variant and a three-seed subsampling check as

robustness controls. Table 6.3 reports the result at $p_t = 0.20$.

| head | cohort | NB before $T$ | NB after $T$ | $\Delta$ | vs treat-none | vs E2 |
|---|---|---|---|---|---|---|
| F2a (FetalCLIP) | uncleaned | $-0.0372 \pm 0.029$ | $\mathbf{+0.0361 \pm 0.019}$ | $+0.073$ | 5/5 | 4/5 |
| F2a (FetalCLIP) | cleaned | $-0.0465 \pm 0.030$ | $\mathbf{+0.0301 \pm 0.027}$ | $+0.077$ | 4/5 | 4/5 |
| E2 (DINOv2 linear probe) | uncleaned | $+0.0149 \pm 0.005$ | $+0.0149$ (no fit) | 0 | 5/5 | — |
| E2 (DINOv2 linear probe) | cleaned | $+0.0163 \pm 0.004$ | $+0.0163$ (no fit) | 0 | 5/5 | — |

Table 6.3: Cross-fold temperature scaling on the auditable-four task at $p_t = 0.20$, five-fold mean $\pm$ SD. 'vs treat-none' counts folds clearing the treat-none reference; 'vs E2' counts folds on which the rescaled FetalCLIP head dominates the E2 reference. The linear probe is already well-calibrated at the threshold, so no temperature is fitted to it.

The fitted temperatures are $T = 0.378 \pm 0.019$ on the uncleaned cohort and $T = 0.385 \pm 0.021$ on the cleaned cohort, stable across the subsampling seeds. The direction is the surprise. A value $T < 1$ *sharpens* the predicted distribution rather than softening it, which is the opposite of the textbook over-confident-transformer correction ($T > 1$). The diagnosis follows from the histograms: the FetalCLIP head's mean predicted probability on the auditable-four slice is 0.30, against a true prevalence of 0.10. It is not over-confident case by case; it is over-calling the base rate by about twenty percentage points, because its binary head was trained on the all-unhealthy prevalence near 0.155 and inherits that bias when restricted to the rarer auditable-four targets. Sharpening is the negative-log-likelihood-optimal response: it pushes the negative-side mass further toward zero and the positive-side mass further toward one, which at the screening thresholds drops false positives faster than true positives. Figure 6.4 shows the correction.

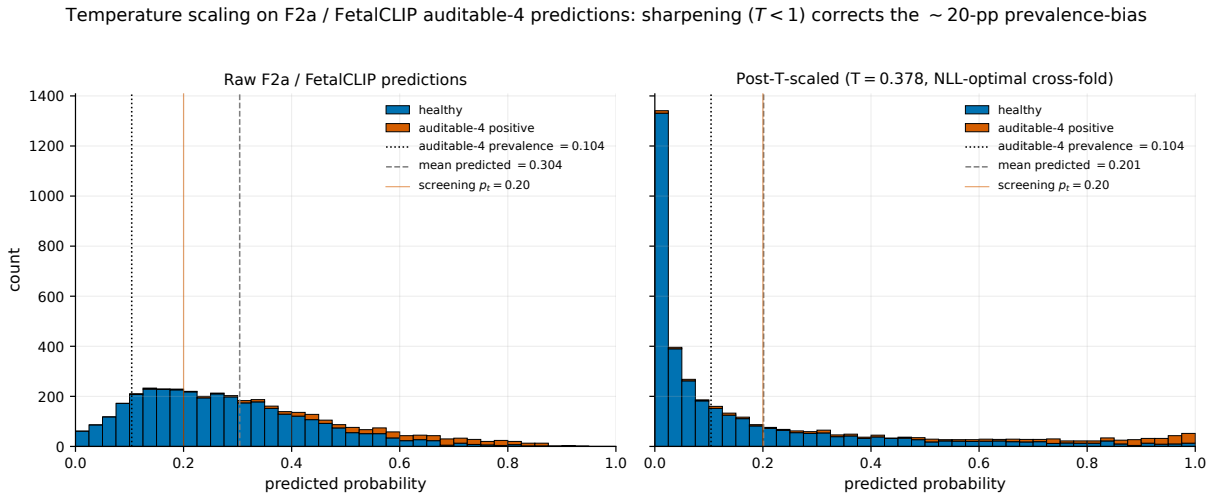


Figure 6.4: Temperature scaling on F2a / FetalCLIP auditable-four predictions, five-fold pooled, $T = 0.378$ (negative-log-likelihood-optimal under the cross-fold protocol). **Left:** raw predicted-probability histogram; the mean prediction (grey dashed) sits at 0.30, about twenty percentage points above the auditable-four prevalence (dotted, 0.10). **Right:** the rescaled distribution, with mean prediction collapsed to 0.20, much closer to the prevalence. The vermilion line marks the screening threshold $p_t = 0.20$. Sharpening ($T < 1$) is the optimal correction for an inherited prevalence bias; the textbook over-confidence remedy is $T > 1$, which would worsen the bias here.

At the five-fold mean, the lift is larger than a closing of the gap — it is a reversal. The rescaled FetalCLIP net benefit at $p_t = 0.20$ exceeds E2's $+0.0149$ by $+0.0212$ on the uncleaned cohort and by $+0.0138$ on the cleaned cohort, both more than one E2 fold-SD (0.0048), and the lead holds at the lower thresholds. The subject-level pooled bias-corrected bootstrap confidence intervals [23]

are non-overlapping at $p_t = 0.20$ on the uncleaned cohort (rescaled F2a $[+0.0265, +0.0457]$ versus E2 $[+0.0048, +0.0241]$), though they overlap on the cleaned cohort and at the lower thresholds.

The per-fold discipline that governs the rest of the chapter tells a stricter story. To displace E2 the rescaled head would have to beat it on all five folds at the threshold; it beats E2 on only four of five on each cohort, and the failing fold differs between cohorts (fold three uncleaned, fold four cleaned), which marks the result as a close call sensitive to the cohort rather than a robust win. On the cleaned cohort the rescaled head even drops one fold below treat-none. The mean-level win is real and substantial, but it does not survive per-fold replication, so the fourth clause is qualified rather than overturned: the linear probe remains the more robust calibration mechanism, and temperature scaling buys the FetalCLIP head a comparable operating point only at the cost of accepting per-fold variance. A pipeline holding both heads could route the high-specificity decision through the linear probe for that robustness while reporting the rescaled FetalCLIP operating point as a complementary calibrated reading, which treats the two backbones as complementary rather than competing.

## 6.7 Reading the model against a sonographer

The decision-curve framework anchors the model to two decision-theoretic references, treat-none and treat-all. Neither is the comparator a clinical reviewer reaches for first; that comparator is the existing UK fetal-anomaly screening service against the same cardiac targets. The relevant model number is the auditable-four sensitivity at fixed specificity 0.95 of $0.435 \pm 0.030$ (Section 5.4, five-seed mean on the test split, FetalCLIP F2a), and the relevant baseline is the most recent national audit [9], which reports per-condition detection rates well above the NHS FASP minimum acceptable floor of 50% [11] across 12,694 diagnoses in NHS England over 2017–2020: TGA 84.9% (95% CI 81.7–87.5), AVSD 69.2% (65.8–72.4), Tetralogy of Fallot 75.4% (72.1–78.4), HLHS 92.7% (90.0–94.8). The model's per-condition sensitivities (slicing-$i$, Section 5.5) sit far below: HLHS 0.600, TGA 0.495, AVSD 0.388, Tetralogy 0.375. The Van Velzen *et al.* 2018 meta-analysis of seven population-based cohorts [8] adds a cross-programme pooled rate of 45.1% (33.5–57.0) for all cardiac defects in unselected populations — at the FASP floor, and reflecting wide variation in training and protocol. Figure 6.5 draws the comparison against the FASP 50% floor.

This is a sense of clinical scale, not a formal benchmark, and the model and published rates sit on different footings: the audit rates are per-condition population detection at the practical specificity of routine reporting (near 0.998), against a lumped model number reported at a deliberately looser 0.95 on a single-centre retrospective hold-out, and a like-for-like read at the deployment specificity is exactly what the cohort's positive count cannot resolve. Even granting the model the favourable end of that mismatch, the conclusion the reviewer should be given is clear in either direction. The lumped sensitivity of 0.435 sits at the Van Velzen cross-programme pooled level (45% [8]) and marginally below the FASP floor, but it is *materially below the modern UK national audit* rates of 69–93% [9] on every cardiac target tested. The binding limitation is not the per-fold variance of Section 6.3 but the cohort's positive count at high specificity together with the discrimination ceiling of the frozen-backbone pipeline at the canonical token budget (Section 5.7).

## 6.8 An alternative explanation: head or backbone?

The dual finding attributes the calibration win to the head architecture. An honest reading has to consider a competing account, because the two candidate causes are confounded in the headline comparison. E2 differs from the FetalCLIP transformer head in two ways at once: it is a linear probe rather than a transformer, and it sits on DINOv2 rather than FetalCLIP. The net-benefit win could therefore be a property of the linear probe — its lack of capacity to inflate mid-range scores, the mechanism this chapter has developed — or it could be a property of the DINOv2

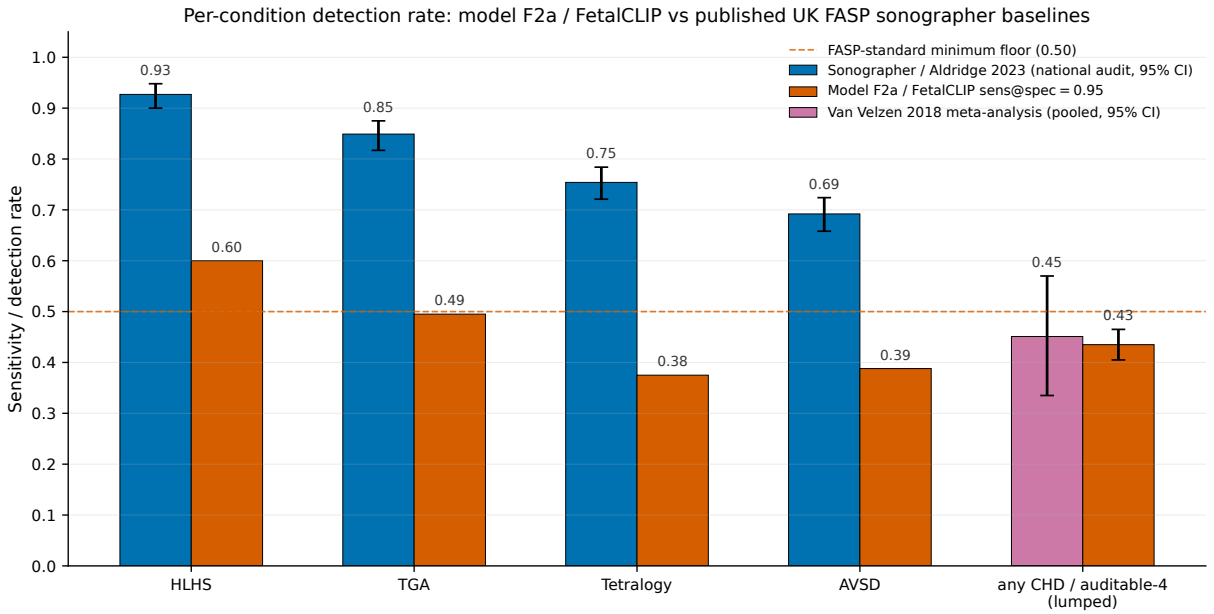


Figure 6.5: Per-condition detection rates: UK national audit [9] sonographer rates (blue, 95% CIs) against the model's per-condition sensitivity at fixed specificity 0.95 (vermilion) on the auditable-four conditions; the lumped comparison sets the Van Velzen 2018 meta-analysis pooled CHD detection (purple) [8] against the model's auditable-four sensitivity. The FASP minimum acceptable floor [11] (0.50) is the horizontal dashed line. The model meets the floor on HLHS only; the comparison is illustrative of clinical scale, with caveats on operating point, lumping, cohort selection, and inter-centre variation discussed in the surrounding text.

representation, whose self-distilled score distribution may simply be more conservative at low base rates than the contrastively pretrained FetalCLIP distribution. Discrimination ranking does not separate these: DINOv2 ranks worse and calibrates better, so backbone and head move together in the headline.

The within-backbone evidence already isolates part of the answer. On DINOv2 the linear probe is the best-calibrated head and the transformer heads are worse (Chapter 4); on FetalCLIP the same ordering holds, with the linear probe again the most threshold-calibrated. The head effect is therefore present on both backbones, which is what the architectural attribution requires. The cleanest test of the residual confound is to put the linear-probe head on the FetalCLIP backbone and ask whether it inherits the linear probe's threshold calibration or the FetalCLIP head's prevalence bias; that isolation, together with a capacity-matched DINOv2-Large control that holds encoder capacity fixed while varying the objective and domain, is the subject of the threats-to-validity analysis in Chapter 7. The claim this chapter defends is the conservative one the evidence already supports: within either backbone, the head is the calibration lever, and across the headline comparison the linear probe on the general-purpose backbone is the configuration that wins the screening decision.

## 6.9 Chapter summary

At the headline screening threshold $p_t = 0.20$ on the auditable-four task, the DINOv2 linear probe returns a net benefit of $+0.0149$ and clears the treat-none reference on all five subject-disjoint folds; both FetalCLIP heads sit below it ($-0.0372$ for the baseline transformer, $-0.0108$ for the four-layer variant), so the model that wins discrimination loses the clinical decision. The Brier–Murphy decomposition reconciles this with the four-layer head's best-in-study aggregate calibration: aggregate calibration and threshold calibration are different objects, and the linear

probe is the threshold-calibration mechanism because it cannot inflate mid-range scores. Cross-fold temperature scaling reverses the net-benefit gap at the five-fold mean ($-0.0372 \rightarrow +0.0361$ uncleaned, $-0.0465 \rightarrow +0.0301$ cleaned) with a sharpening temperature $T \approx 0.38$ that corrects an inherited prevalence bias rather than ordinary over-confidence, but it beats E2 on only four of five folds and so qualifies the architectural mechanism rather than overturning it. Read against published sonographer rates, the model's auditable-four sensitivity at fixed specificity 0.95 (0.435) sits at the cross-programme meta-analytic level and the FASP floor but materially below the modern UK national audit, which is the comparison a clinical reviewer should be given. The head-versus-backbone confound is named and partly isolated here; Chapter 7 takes up the remaining controls, the threats to validity, and the open frontier.

# Chapter 7

# Discussion, Ethics, and Future Work

The preceding three chapters did the measuring. This one does the thinking: what the dual finding means once the tables are set aside, why the scaling experiment landed where it did, which of the conclusions a sceptical examiner could legitimately attack, and what an honest reading owes the reader in the way of ethics and unfinished business. The headline numbers belong to Chapter 8. Here the question is what they are worth.

## 7.1 What the dual finding means

The result the dissertation defends is the dual finding stated as the spine of Chapter 1: the domain-specialist backbone wins discrimination, the simplest head on the general-purpose backbone wins the clinical decision, the head architecture is the calibration lever, and temperature scaling halves but does not close the gap. Read against the prevailing intuition in medical imaging, this is a mild heresy. The field default is that a domain-specialist representation strictly dominates a general-purpose one: a backbone pretrained on fetal ultrasound should beat a backbone pretrained on natural images, and the only honest question is by how much. On the discrimination axis that default holds here, and cleanly. Swapping to FetalCLIP lifts the auditable-four AUROC from $0.818 \pm 0.022$ to $0.911 \pm 0.022$ across five subject-disjoint folds, a gain of roughly four fold-standard-deviations on a task where the two backbones share an interface, a head family, a recipe, and an evaluation grid (Chapter 5). If the question were *which representation ranks sick above healthy more often*, the answer would be FetalCLIP and the dissertation would be a paragraph long.

The decision axis breaks the default. At the screening threshold $p_t = 0.20$ the DINOv2 linear probe returns a per-fold net benefit of $+0.0149$ and clears the treat-none reference on all five folds, while the higher-AUROC FetalCLIP transformer head sits at $-0.0372$ and clears treat-none on none of them (Chapter 6). The better discriminator is the worse decision tool at the operating point a screening service would actually adopt. That is not a rounding artefact and it is not a quirk of one threshold: the ranking is stable across $p_t \in \{0.10, 0.15, 0.20\}$, and it holds on both the raw and the labelling-cleaned cohorts.

The mechanism that reconciles the two axes is architectural, and it sits in the head rather than the backbone — the mechanism developed in Section 6.5. The Brier–Murphy decomposition makes this precise (Section 6.4): the head with the best aggregate calibration in the whole study loses the threshold decision, because aggregate calibration averages over the full $[0, 1]$ interval while net benefit cares only about the narrow band around $p_t$. Temperature scaling, the obvious post-hoc fix, qualifies rather than overturns this. A sharpening temperature ($T \approx 0.38$) reverses the FetalCLIP-versus-DINOv2 net-benefit gap at the five-fold mean but beats the linear probe on only four of five folds, failing the per-fold replication bar the rest of the analysis holds itself to (Section 6.6).

The synthesis is therefore narrow and load-bearing at once. A domain-specialist backbone is

neither necessary for the best screening decision nor sufficient to secure it. The lever that decides the clinical contest is the head, and the simplest head wins it. For a practitioner choosing a model for a fixed-threshold screening task, the actionable reading is that representation quality and decision quality are separable, that the second is bought with calibration rather than discrimination, and that a one-parameter rescaling buys a comparable operating point only at the price of fold-to-fold robustness.

## 7.2 Why scaling did not lift clinical utility

The scaling experiment was meant to be the obvious escape hatch. If the FetalCLIP head's net-benefit deficit were a starvation effect, then feeding the model more of each study should close it. Chapter 5 reports what happened instead: across the budgets tested, up to three times the canonical two-hundred-token allowance, discrimination edges upward only modestly and within fold variance, while the clinical net-benefit ceiling does not move at all. The decision-curve result stays negative on every fold at every increment. The discrimination curve has a faint positive slope; the utility curve is flat against the screening reference.

Two properties of the training signal explain why more tokens do not help where it matters. The first is a gradient-budget problem. The classifier is trained at the subject level: the loss is computed once per subject from the aggregated prediction, so every per-clip parameter inside a hierarchical head receives a gradient that is diluted across all the clips that fed the aggregate. A per-clip encoder trained under a subject-level loss sees roughly twenty times less learning signal per parameter than the same encoder would under a clip-level loss, which is why the four head-side hypotheses that could have lifted the ceiling were falsified in turn (Section 5.7): head capacity, training recipe, hierarchical clip aggregation, and view-aware clip selection each fail to move the utility ceiling, isolating the binding constraint to the representation itself under a data-starved objective. The second property is a sampling problem. Frames are drawn across each clip with `np.linspace`, which at the native frame rate places successive samples about one fetal cardiac cycle apart. The head therefore sees structural snapshots at uncorrelated cardiac phases rather than the motion of a beating heart. Adding more such snapshots adds more views of the same structure, not the temporal signal that distinguishes, say, a coarctation from a normal arch on a dynamic sweep. More budget buys more of the wrong axis of information.

The honest scope of the null is the part that matters for the reader. The ceiling sits at three times the canonical budget, not at any budget. The experiment bounds the response of discrimination and the invariance of utility over the range it tested; it does not license the stronger claim that no budget would ever break through. That stronger question is well-posed and deliberately left to the open frontier of Section 7.6, where it carries a pre-registered prediction rather than a hand-wave.

## 7.3 Threats to validity and alternative explanations

A finding that contradicts the field default earns extra scrutiny, and the candid accounting of what could undermine it is where most of the dissertation's claim to rigour is cashed. The threats fall into two groups: alternative explanations that could re-attribute the dual finding to a different cause, and limitations that bound how far it generalises.

**The head-versus-backbone confound.** Two things change at once in the headline comparison, and the win has to be charged to one of them. The DINOv2 linear probe differs from the FetalCLIP transformer head both in its head architecture (linear versus transformer) and in its backbone (self-distilled DINOv2 versus contrastive FetalCLIP). The net-benefit win could therefore be a property of the linear probe, the mechanism this dissertation develops, or it could be a property of the DINOv2 representation, whose self-distilled score distribution may simply be

more conservative at a low base rate than the FetalCLIP distribution. Discrimination ranking cannot separate the two, because DINOv2 ranks worse and calibrates better, so backbone and head move together. The within-backbone evidence settles the attribution, and it is already in hand: the linear probe is the most threshold-calibrated head on *both* backbones (Chapters 4 and 6), which is precisely what an architectural account requires and what a pure-backbone account does not predict — a representation effect would not hold the head ranking fixed as the representation changes. The claim the evidence already supports, and the one the dual finding rests on, is that within either backbone the head is the calibration lever. A cross-backbone check sharpens the picture further without being load-bearing for it: putting the linear-probe head on the FetalCLIP backbone (Section 6.8) asks whether it inherits the probe's threshold calibration or the FetalCLIP head's prevalence bias, and a confirming result would close the last gap between the two backbones rather than establish the mechanism, which the within-backbone evidence has already done.

**The backbone-capacity confound.** Encoder capacity is the one axis the swap does not hold fixed. The comparison is matched at the 768-dimensional embedding interface — the head sees an identically-shaped input from either backbone — but DINOv2 is a ViT-B/14 with about 86 million parameters while FetalCLIP's image tower is a ViT-L/14 with about 300 million. The discrimination lift could therefore be a domain effect, a capacity effect, or both, and the dissertation names this openly rather than assuming the domain reading. The evidence in hand already constrains the question: the gain survives across five subject-disjoint folds at a shared interface, head family, recipe, and evaluation grid, so whatever its source it is a stable property of the representation rather than an artefact of the comparison. What the present evidence cannot do is apportion that stable gain between domain and size, and a capacity-matched control is the planned check that would. A DINOv2 ViT-L arm holds encoder capacity at the FetalCLIP scale while varying only the objective and the pretraining domain, so the gap between it and FetalCLIP would isolate the contribution of domain-specialist pretraining net of size; the run is staged on the project compute and is best read as forthcoming confirmation — it would sharpen the domain-versus-capacity reading of an already-established lift, not decide whether the lift exists. The clinical half of the dual finding does not turn on the outcome at all: the calibration mechanism is a head property demonstrated within both backbones, and it is untouched by how the discrimination gap is finally apportioned between domain and capacity.

**Antenatal-suspicion labels.** The nine condition columns record antenatal suspicion rather than postnatal confirmation, with the single exception of the coarctation slice cleaned in Section 5.6. For the other eight conditions the healthy class plausibly contains an unknown fraction of subjects with undiagnosed disease. This bounds the *absolute* interpretation of every AUROC and net-benefit estimate upward: the comparisons across architectures are internally valid, because every model sees the same labels, but the absolute values are upper bounds on what postnatally-confirmed labels would yield. The coarctation cleanup is a small proof that the direction of the dual finding survives a label-quality intervention, since it lifts that condition's five-fold AUROC from about 0.85 to 0.879 with no change to any model and no change to the headline ordering.

**Single-centre cohort and per-condition power.** iFIND is one large Imperial-affiliated centre, so external validity to other scanners, sonographers, and populations is an open empirical question rather than a settled one. The cohort is moreover not stratified by maternal demographics, scanner model, or operator, so subgroup performance is unknown — a threat to external validity treated as an ethics matter in Section 7.4. The per-condition picture is statistically thin by construction: eight of the nine conditions have fewer than twenty positives, the bias-corrected-and-accelerated confidence intervals [23] are wide, and the per-condition deltas between backbones

overlap. That is the reason the dissertation leads every inferential claim with the lumped auditable-four metric and treats the per-condition cells as descriptive (Section 5.5). It is also why the auditable-four estimand, rather than the all-nine or the per-condition surface, is the clinical headline: it is the slice with the prevalence and the positive count to load-bear an operating-point decision, and it is the slice that maps onto the UK screening targets a service would actually audit.

**Operating-point and estimand sensitivity.** The dual finding is articulated at $p_t = 0.20$ with the lower thresholds as corroboration; a materially different operating point, such as a diagnostic rather than screening regime, could re-rank the architectures, because decision-curve net benefit at a fixed $p_t$ is by design sensitive to the threshold. The ranking is stable across the screening band, but the claim is a screening claim and is scoped as one. The choice of the lumped auditable-four estimand over the all-nine estimand is likewise a deliberate decision with a stated rationale rather than a free parameter: all-nine is reported as the secondary surface, the per-condition cells as the tertiary, and the lumped four as the load-bearing one.

## 7.4 Ethics and responsible AI

The dissertation re-analyses a de-identified retrospective research cohort and trains no model intended for deployment. Its claims are about evaluation, not about a fielded tool, and the ethical framing follows from that boundary. The work falls under the iFIND project's existing institutional ethics approval, was carried out on Imperial-managed infrastructure restricted to authorised collaborators, and involved no new data collection or human-participant intervention. The Department of Computing ethics checklist items it engages are those covering the processing of personal data, the use of an existing dataset containing personal information, medical research, and overlap with NHS clinical services; the detail sits in the Declarations.

A handful of responsible-AI points deserve to be made in the body rather than the appendix. There is, first, the regulatory boundary. The outputs are clinical-decision-support research, not a medical device, and under that framing they do not engage the device-regulation regime; any prospective deployment of the two-head structure the dual finding suggests would require its own conformity assessment, a clinical-performance study, and a post-market surveillance plan, none of which this work supplies or pretends to.

It is worth being concrete, and explicitly speculative, about the shape such a deployment might take, precisely so that the caveats above attach to something specific. A forward-looking design — forward-looking, not a validated recommendation — would pair two heads: the DINOv2 mean-pool linear probe (E2) acting as a high-specificity screening gate, and the post-temperature-scaled FetalCLIP F2a head supplying a complementary calibrated probability estimate alongside it. The intended setting is decision support at a tertiary-referral screening gate, where a clinician retains the decision, and emphatically not autonomous screening. None of this is licensed by the present evidence: it would still demand the conformity assessment, clinical-performance study, and post-market surveillance named above before any patient ever saw its output.

The principal anti-harm contribution is methodological. A practitioner who selected a screening model on discrimination alone would prefer the FetalCLIP head, with its 0.911 auditable-four AUROC over the linear probe's 0.818; the decision-curve analysis at the screening threshold inverts that preference. Making that inversion concrete, on a real cohort with a clinically-grounded operating point, is the safeguard the dissertation offers against a real and common selection error. The error matters because of the asymmetric error budget set out in Chapter 1, which is exactly why the analysis is anchored at a high-specificity operating point [11, 12] rather than at the threshold that maximises some symmetric accuracy.

The last point is honesty about where the model stands against the service it would notionally assist. The model's auditable-four sensitivity at a fixed specificity of 0.95 is 0.435, which sits at

the cross-programme meta-analytic level and at the NHS screening floor but materially below the modern UK national audit, where detection of the four cardiac targets runs from about 69% to 93% (Section 6.7). The dissertation states this plainly: the model is not at modern UK-audit detection rates, and nothing in the dual finding should be read as a claim that it is ready to help a sonographer today. Two further limitations are named for the same reason. The iFIND cohort's demographic strata are not analysed, so sub-population disparities in discrimination or calibration are unmeasured and would have to be characterised before any deployment. And because the work uses only frozen backbones with cached embeddings, its compute footprint is confined to head-training, with no fresh self-supervised pretraining; the substantial pretraining cost of the two backbones is amortised across every experiment rather than repeated. The use of generative-AI tooling in preparing the dissertation is disclosed in full in the Declarations.

## 7.5 Reporting against TRIPOD+AI

The relevant reporting standard for a machine-learning clinical-prediction-model study is the TRIPOD+AI statement [10], and the dissertation engages it item by item, with the full checklist in Appendix A. The load-bearing items are met. The cohort, eligibility criteria, and subject-disjoint development and evaluation splits are specified in Chapter 3; the frozen-backbone-plus-head predictors and the outcome definitions, including the auditable-four lumping, are specified in Chapters 4 and 5; the discrimination assessment reports AUROC with paired five-fold bias-corrected-and-accelerated confidence intervals [23], and the calibration assessment reports the Brier–Murphy decomposition [22], reliability diagrams, and a temperature-scaling intervention [5]. The clinical-utility evidence the standard expects under interpretation is the per-fold decision-curve analysis at the pre-specified screening grid, reported with per-fold positive-count discipline rather than a pooled estimand [3].

The principal unmet item is external validation. The evaluation is internal throughout: a five-fold subject-disjoint replication on iFIND plus a locked test split never used for model selection. No external-cohort evaluation has been performed, so the dissertation's claims are bounded by single-cohort generalisability until one is. The PROMETHEUS cohort [13] is the natural external dataset, an external evaluation against it is pre-staged subject to data access, and Section 7.6 names it as the single most consequential future-work item. One further item is only partially met: a formal prospective sample-size calculation [25] is not performed, because the cohort is a fixed retrospective resource rather than a powered study, and the compensating discipline is the wide per-condition confidence intervals reported wherever positive counts are small.

## 7.6 Future work

The cleanest open question is the one the scaling null was careful not to answer. The dissertation bounds the discrimination response and the utility invariance up to three times the canonical budget; the pre-registered frontier is the asymptotic regime at roughly twenty times baseline, a per-subject token budget of order $nf = 60 \times nc = 60$, where the model finally sees a substantial fraction of each study. The experiment is specified with a decision rule fixed in advance, which is what keeps it from becoming a fishing expedition. If discrimination lifts there, that is a contribution: it locates the budget at which a domain-specialist representation starts to pay off. If it saturates, the saturation is itself the answer, because it pins the binding constraint to the representation under a subject-level signal and converts 'we did not find a lift' into 'this is the gradient, and exponentially more data is needed to move it'. The frontier carries a sharper prediction still: the two structure-dominated conditions, right aortic arch and coarctation, are expected to sit near chance, mean AUROC at or below 0.4, if the head is bound by scanning style rather than image signal, and to rise if it is image-signal-bound. A clear test either way

is a result. When the asymptotic run lands it will be reported with per-fold decision-curve net benefit and sensitivity at fixed specificity, not with AUROC alone, because the whole argument of this dissertation is that AUROC is the discrimination check and not the headline.

Further directions follow from the mechanisms of Section 7.2. The sampling problem invites a multi-window correction: cutting each clip into several short contiguous native-frame-rate bursts, each a fixed-cost training example, so the head sees the motion of a cardiac cycle rather than the structural snapshots that `np.linspace` delivers. The gradient-budget problem invites a clip-level pretext task that learns per-clip features under a clip-level loss before a subject head is attached on top, sidestepping the twenty-fold dilution that starves a per-clip encoder trained end-to-end. And the per-condition thinness invites condition-specific heads, each attending to the embedding sequence through the diagnostic view that condition actually uses, in place of a single shared head that averages over nine different view mechanisms; the caution there is the project's own discipline of evaluating one method well rather than many shallowly, so a per-condition head set is worth building only once the shared-head baseline is fully characterised.

Above all of these sits external validation. The model has been replicated across folds and locked against a held-out split, but it has not left its cohort, and that is the gap between an internally-consistent result and a clinically-transportable one. Taking the iFIND-trained linear probe and FetalCLIP heads unchanged, freezing them, and evaluating on PROMETHEUS [13] with the same auditable-four AUROC, per-condition AUROC, and per-fold decision-curve discipline used internally is the experiment that would turn the dual finding from a claim about one centre's data into a claim about the method. It is pre-staged and waiting on data access. Running alongside external validation, and a precondition for any prospective use rather than an optional extra, is a subgroup-fairness evaluation: a demographic-stratified re-evaluation of discrimination and calibration across maternal demographics, scanner model, and operator, which the present single-cohort analysis leaves entirely unmeasured (Section 7.3). Until both run, the dissertation's contribution is a controlled, calibration-aware backbone swap on a real cohort that separates discrimination from clinical utility and locates the calibration lever in the head; what it is not, yet, is an externally-validated one, and saying so is part of the result.

# Chapter 8

# Conclusion

## 8.1 What the dissertation establishes

This dissertation asked a deliberately narrow question and answered it with the full apparatus a clinical-prediction study deserves. On the iFIND fetal-cardiac cohort of 4,128 subjects, the frozen image backbone that produces per-frame embeddings was swapped from a general-purpose self-distilled model to a domain-specialist vision–language model, with the classifier head, training recipe, subject-disjoint splits, and evaluation grid all held constant. Because both backbones emit a 768-dimensional embedding, the exchange is genuinely like-for-like. Everything that moved in the results moved because of the backbone, the head, or the calibration step, and not because the measurement surface had changed underneath them. The headline is the dual finding of Chapter 1: the domain-specialist backbone wins discrimination, the simplest head on the general-purpose backbone wins the clinical decision, and the head is the calibration lever between them.

The work behind that one finding resolves into five completed contributions.

**A controlled, calibration-aware backbone swap.** The swap is the spine of the dissertation, and it produces the dual finding in full. On the auditable-four task — the four UK screening targets lumped into a single binary label — FetalCLIP lifts five-fold AUROC from $0.818 \pm 0.022$ to $0.911 \pm 0.022$, a gain of roughly four fold-standard-deviations with non-overlapping paired-bootstrap intervals. At the screening threshold $p_t = 0.20$ the ranking inverts: the DINOv2 linear probe clears the treat-none reference on five of five folds with a per-fold mean net benefit of $+0.0149$, while the higher-AUROC FetalCLIP transformer head sits below treat-none on zero of five, at $-0.0372$, across the whole screening band $p_t \in \{0.10, 0.15, 0.20\}$. The model that ranks cases better is the model that helps less at the threshold a service would adopt, and the finding holds on both the all-nine and the auditable-four tasks and on both the raw and the postnatal-confirmation-cleaned cohorts.

**A per-fold decision-curve methodology, established by self-correction.** A first-pass scaling result compared a pooled-on-test-split net-benefit estimate against a per-fold estimate on the same predictions and manufactured a spurious 'scaling beats the baseline' headline. An internal review caught the mismatch before it left the project. The claim was withdrawn, and per-fold net benefit was adopted as the canonical estimand for $k$-fold settings in which each fold trains on a distinct subset and calibrates on its own validation split [3]. The two estimands can disagree materially on identical predictions, and the pooled one is defensible only under exchangeability assumptions that subject-disjoint cross-validation violates by construction. The discipline that followed is what every clinical claim in the dissertation is held to: per-fold net benefit, BCa rather than percentile bootstrap intervals [23], and a pre-registered bar for elevating one model above another.

**A four-way falsification of head-side scaling hypotheses.** The flat clinical-utility ceiling under token scaling was tested against four candidate causes and survived all of them. Head capacity, training-recipe tuning, hierarchical clip aggregation, and cardiac-view-aware clip selection were each shown not to be the binding constraint. Only added head depth moved anything, and it improved aggregate calibration on the FetalCLIP backbone without transferring that improvement to the screening threshold. Four nulls and one qualified positive isolate the constraint to the image representation under a data-starved subject-level training signal, which is a sharper diagnosis than the flat curve alone would license.

**A temperature-scaling qualification.** Cross-fold temperature scaling [5] lifts the FetalCLIP head to or above the linear probe at the five-fold mean, but it fails the per-fold replication bar at four of five folds rather than five, so it cannot be relied on at the operating point. The fitted temperature sharpens rather than softens the scores, which diagnoses a prevalence-inheritance bias — the FetalCLIP head over-calls the auditable-four base rate because it was trained on a higher all-nine prevalence — rather than ordinary per-instance over-confidence. The architectural-mechanism clause of the dual finding therefore stands; the temperature-scaling clause is qualified, not overturned.

**A postnatal-confirmation cohort cleanup.** A confirmation flag encoded in the iFIND filename field was decoded to restrict the coarctation-of-the-aorta slice — the condition with the weakest antenatal positive predictive value — to postnatally-confirmed positives. The restriction lifts the five-fold coarctation AUROC from roughly 0.85 to 0.879 with no retraining and no change to any model, and the auditable-four dual finding survives the cleanup on both cohorts.

## 8.2 What the result means for the field

The prevailing assumption in medical imaging is that a domain-specialist representation, pre-trained on in-domain data, strictly dominates a general-purpose one. This dissertation does not contradict that assumption on its own terms, and it does not need to: FetalCLIP does win discrimination, cleanly and reproducibly. The point is that discrimination is not the whole contest. When the task is low-prevalence screening and the operating point is fixed in advance, the domain-specialist backbone wins the ranking and loses the decision, and the lever that decides the decision is the head architecture rather than the backbone. The mechanism developed in Section 6.5 settles why: it is the head, not the representation, that governs where the probability mass sits relative to the screening threshold.

The result generalises past the particular pair of backbones that produced it. A better representation buys you a better ROC curve. It does not buy you a better clinical decision unless the head that reads it is calibrated at the point of use. For a field that selects backbones on AUROC and treats the head as an afterthought, the ordering is exactly backwards: the head is where the screening decision is won or lost.

## 8.3 The open frontier

One question is left open, and it is open by design rather than by omission. The scaling experiment bounded the response of discrimination and clinical utility across budgets up to three times the canonical token allowance. Within that range the answer is settled: discrimination edges up modestly and within fold variance, the net-benefit ceiling does not move at all, and four head-side mechanisms are falsified. What that bounded experiment cannot reach is the behaviour at an order of magnitude more data. The asymptotic regime, roughly twenty times the budget tested here, is the named next step, and it carries a pre-registered prediction: if the residual ceiling on the hardest conditions is scanning-style-bound rather than image-signal-bound,

their mean AUROC should stay near chance even at that scale, whereas an image-signal-bound ceiling should lift. The experiment and the decision rule for reading its outcome are both fixed in advance, so the asymptotic run can only confirm or refute the saturation reading, not be tuned to fit it; only the compute remains. The continuation is set out in Chapter 7.

External validation against an independent cohort sits alongside it as the other honest open item [10]. The iFIND cohort is single-centre, and the dissertation says plainly where its claims stop. What it offers in their place is a method for asking the question correctly.

## 8.4 The methodological takeaway

If this dissertation leaves one durable instruction for low-prevalence screening studies, it is in how to judge a model rather than in which model to pick. Chapter 7 makes the case that discrimination is a check and not a verdict; the instruction that follows from it is a single procedure. Evaluate on per-fold net benefit at the pre-specified operating point, with confidence intervals that hold up at the discrete thresholds screening turns on, and report the result against an honest clinical baseline. A study that does only this much would not have selected the auditable-four headline of 0.911 over 0.818 and deployed the model that loses net benefit on every fold. That is the question this dissertation was built to answer, and the procedure is the answer that outlasts the particular backbones.

# References


[1] Denise van der Linde et al. 'Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis'. In: *Journal of the American College of Cardiology* 58.21 (2011), pp. 2241–2247. DOI: 10.1016/j.jacc.2011.08.025.

[2] Steve Halligan, Douglas G. Altman and Susan Mallett. 'Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach'. In: *European Radiology* 25.4 (2015), pp. 932–939. DOI: 10.1007/s00330-014-3487-0.

[3] Andrew J. Vickers and Elena B. Elkin. 'Decision Curve Analysis: A Novel Method for Evaluating Prediction Models'. In: *Medical Decision Making* 26.6 (2006), pp. 565–574. DOI: 10.1177/0272989X06295361.

[4] Andrew J. Vickers, Ben van Calster and Ewout W. Steyerberg. 'A simple, step-by-step guide to interpreting decision curve analysis'. In: *Diagnostic and Prognostic Research* 3 (2019), p. 18. DOI: 10.1186/s41512-019-0064-7.

[5] Chuan Guo et al. 'On Calibration of Modern Neural Networks'. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1321–1330. arXiv: 1706.04599. URL: https://proceedings.mlr.press/v70/guo17a.html.

[6] Maxime Oquab et al. 'DINOv2: Learning Robust Visual Features without Supervision'. In: *Transactions on Machine Learning Research* (Jan. 2024). arXiv: 2304.07193 [cs.CV]. URL: https://arxiv.org/abs/2304.07193.

[7] Fadillah Maani et al. 'FetalCLIP: A Visual-Language Foundation Model for Fetal Ultrasound Image Analysis'. In: *arXiv preprint* (2025). arXiv: 2502.14807 [eess.IV]. URL: https://arxiv.org/abs/2502.14807.

[8] Christine L. van Velzen et al. 'Systematic review and meta-analysis of the performance of second-trimester screening for prenatal detection of congenital heart defects'. In: *International Journal of Gynaecology & Obstetrics* 140.2 (2018), pp. 137–145. DOI: 10.1002/ijgo.12373.

[9] Natalie Aldridge et al. 'Detection rates of a national fetal anomaly screening programme: A national cohort study'. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 130.1 (2023), pp. 51–58. DOI: 10.1111/1471-0528.17287.

[10] Gary S. Collins et al. 'TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods'. In: *BMJ* 385 (2024), e078378. DOI: 10.1136/bmj-2023-078378. URL: https://www.bmj.com/content/385/bmj-2023-078378.

[11] NHS Fetal Anomaly Screening Programme. *NHS Fetal Anomaly Screening Programme: programme handbook*. GOV.UK. Updated 22 April 2026; current guidance for the NHS England fetal anomaly screening (FASP) 20-week scan, including cardiac screening targets for TGA, AVSD, Tetralogy of Fallot, HLHS, and CoA. Cardiac detection-rate data last updated 15 May 2025. Apr. 2026. URL: https://www.gov.uk/government/publications/fetal-anomaly-screening-programme-handbook (visited on 28/05/2026).

[12] Julene S. Carvalho et al. 'ISUOG Practice Guidelines (updated): fetal cardiac screening'. In: *Ultrasound in Obstetrics & Gynecology* 61.6 (June 2023), pp. 788–803. DOI: 10.1002/uog.26224.

[13] Thomas G. Day et al. 'Artificial Intelligence to Assist in the Screening Fetal Anomaly Ultrasound Scan (PROMETHEUS): A Randomised Controlled Trial'. In: *medRxiv* (2024). Preprint, posted 2024-05-25. DOI: 10.1101/2024.05.23.24307329. URL: https://www.medrxiv.org/content/10.1101/2024.05.23.24307329v1.

[14] Ashish Vaswani et al. 'Attention Is All You Need'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 6000–6010. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[15] Alexey Dosovitskiy et al. 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'. In: *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[16] Mathilde Caron et al. 'Emerging Properties in Self-Supervised Vision Transformers'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 9650–9660. arXiv: 2104.14294 [cs.CV]. URL: https://arxiv.org/abs/2104.14294.

[17] Alec Radford et al. 'Learning Transferable Visual Models From Natural Language Supervision'. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763. arXiv: 2103.00020. URL: https://proceedings.mlr.press/v139/radford21a.html.

[18] Michael Moor et al. 'Foundation models for generalist medical artificial intelligence'. In: *Nature* 616.7956 (2023), pp. 259–265. DOI: 10.1038/s41586-023-05881-4.

[19] Sheng Zhang et al. 'BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs'. In: *arXiv preprint* (2023). arXiv: 2303.00915 [cs.CV]. URL: https://arxiv.org/abs/2303.00915.

[20] Ming Y. Lu et al. 'A visual-language foundation model for computational pathology'. In: *Nature Medicine* 30 (2024), pp. 863–874. DOI: 10.1038/s41591-024-02856-4.

[21] Jianbo Jiao et al. 'USFM: A Universal Ultrasound Foundation Model Generalized to Tasks and Organs towards Label Efficient Image Analysis'. In: *Medical Image Analysis* 96 (2024), p. 103202. DOI: 10.1016/j.media.2024.103202. arXiv: 2401.00153 [eess.IV].

[22] Allan H. Murphy. 'A New Vector Partition of the Probability Score'. In: *Journal of Applied Meteorology* 12.4 (1973), pp. 595–600. DOI: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

[23] Bradley Efron. 'Better Bootstrap Confidence Intervals'. In: *Journal of the American Statistical Association* 82.397 (1987), pp. 171–185. DOI: 10.1080/01621459.1987.10478410.

[24] Ben Van Calster et al. 'Calibration: the Achilles heel of predictive analytics'. In: *BMC Medicine* 17 (2019), p. 230. DOI: 10.1186/s12916-019-1466-7.

[25] Richard D. Riley et al. 'Calculating the sample size required for developing a clinical prediction model'. In: *BMJ* 368 (2020), p. m441. DOI: 10.1136/bmj.m441.

# Declarations

## Use of Generative AI

**Tool:** Claude Code (CLI), powered by Anthropic Claude (Opus 4.7 and 4.8)
**Publisher:** Anthropic
**URL:** https://claude.com/claude-code
**Period of use:** 2026-04 through 2026-06 (the duration of this project)

**Scope of use.** Claude Code was used in three categorically distinct roles throughout this project, each with author oversight and auditable artefacts retained for academic-integrity verification:

1. **Experimental orchestration.** Dispatching parallel head-training jobs across the lab's GPU resources, running bias-corrected-and-accelerated bootstrap aggregations on saved per-fold predictions, regenerating decision-curve and reliability figures from canonical CSV outputs, maintaining a source-of-truth directory (`docs/source-of-truth/`) that records project state, hypothesis tree, and lessons learned across sessions. The agent did not invent experimental decisions: every architecture, hyperparameter, evaluation slice, and revisability criterion was pre-specified by the author and documented in the source-of-truth materials before execution.

2. **Initial prose drafting.** The Phase 1 initial drafts of Chapters 1, 2, 3, 4, 7, and 8, plus the abstract, plus structural prose in Chapters 5 and 6, plus the ethics and sustainability declarations, were produced by Claude subagents against author-written prompts that specified section organisation, required reading lists, citation policies, and page budgets. Subagents returned LaTeX content together with verification reports documenting what was written, which numbers were drawn from existing experiment files, and which choices were flagged for author review. The author read every chapter, validated every load-bearing number against the underlying experiment outputs, re-wrote sentences that did not match the intended argumentative cadence, and made every final editorial decision on what to retain.

3. **Reading and verification aid.** The agent was used to cross-check numbers across chapters, surface internal inconsistencies, and verify that bibliographic entries matched the form required by the dissertation style. The `scripts/session_autocommit.sh` infrastructure preserves a per-session checkpoint commit trail.

**What was *not* done by the agent.** The research-question framing, the dual-finding interpretation, the methodology audit walked back during the project, the per-fold DCA canon, the auditable-four T-scaling qualification, the pre-registered revisability criteria in Chapter 6, the clinical-deployment recipe in Chapter 7, the threats-to-validity inventory, and the architectural identification of the linear probe as the calibration mechanism are all the author's own contributions, developed in consultation with the supervisor (Professor Bernhard Kainz) and clinical authority (Dr Tom Day). The agent did not propose research questions, did not interpret results, and did not make methodological calls that were not first specified by the author.

**Audit trail.** Full session prompts and per-session orchestrator logs used to produce this dissertation are retained locally under `~/.claude/projects/`, alongside the `scripts/session_autocommit.sh` per-turn commit checkpoints on the project's feature branch. These artefacts are available on request for academic-integrity verification.

I confirm that no AI-generated content has been presented as my own original thinking or as factually unverified output. Where AI-generated prose contributed to chapter drafts, the final wording has been read, validated against underlying source data and supervisor guidance, and edited by the author prior to inclusion.

## Ethical Considerations

**Study design.** This dissertation is a retrospective, computational study using de-identified fetal ultrasound video acquired through the Imperial College London iFIND research programme, of which the supervisor (Professor Bernhard Kainz) is a co-investigator. No new human-participant intervention, recruitment, or data collection was carried out for this dissertation; all data were de-identified at source under the iFIND project's existing institutional ethics approval and data-sharing governance, and access was restricted to authorised collaborators on Imperial-managed infrastructure for the duration of the project. Against the Department of Computing ethics checklist (`reference/ethics-checklist-template.md`), items 6 (processing of personal data), 7 (use of existing datasets containing personal information), 26 (medical research), and 27 (overlap with NHS clinical services) apply; the remaining checklist items do not.

**Regulatory framing.** The project's outputs are framed as clinical-decision-support *research* rather than a deployed medical device, and therefore do not engage the UK Medicines and Healthcare products Regulatory Agency (MHRA) medical-device regulatory regime in their current form. Any prospective clinical deployment of the heads and procedures reported here — on the iFIND, PROMETHEUS, or any other prospective cohort — would require separate regulatory consideration and is explicitly out of scope.

**Asymmetric clinical harms.** Fetal congenital-heart-disease findings carry direct implications for downstream patient-management decisions. The dissertation's choice of operating point (sensitivity at specificity 0.95) reflects the clinically asymmetric error budget: the dominant harm from a false-positive flag is anxiety to the family and an additional specialist scan, whereas the dominant harm from a false-negative miss is undiagnosed disease at birth with materially worse paediatric outcomes. The dual finding reported in this dissertation (Tier 1 wins clinical utility at the screening-relevant threshold even where Tier 2 wins on a discrimination-only metric) is explicitly framed to flag the risk of model selection by AUROC alone in a clinical context, and is the dissertation's principal anti-harm methodological contribution.

**Fairness and subgroup analysis.** The iFIND cohort's demographic stratification (maternal age, ethnicity, gestational-age distribution, scanner-model heterogeneity, sonographer skill profile) is not formally analysed in this dissertation; subgroup fairness across these dimensions is therefore an unaddressed limitation that should be evaluated before any prospective clinical use. This is acknowledged as an explicit limitation in Section 7.3 (threats to validity) and as future work in Section 7.6.

**Sustainability.** Computational sustainability is discussed in the Sustainability declaration immediately below; in summary, the project amortises pretraining compute by using frozen backbones (DINOv2 and FetalCLIP) across all downstream experiments, caches per-subject embeddings to disk to avoid redundant forward passes, budgets hyperparameter sweeps ahead of time, and prunes search aggressively when fold-zero results indicate no signal.

**Audit trail.** The author's use of generative-AI assistance during this project is declared transparently in the Use of Generative AI section immediately above. The audit trail (per-session

prompts, per-turn commit checkpoints, source-of-truth documentation) is retained locally and available on request for academic-integrity verification.

To the best of the author's knowledge, no additional ethical issues beyond those addressed by the iFIND project's existing institutional ethics approval and data-sharing agreement remain for this dissertation.

## Sustainability

The computational experiments reported in this dissertation were run on shared GPU resources operated by the Imperial College Department of Computing and on a collaborator-provided four-GPU workstation. Frozen backbones (DINOv2 and FetalCLIP) were used wherever possible to amortise pretraining cost across all downstream experiments, and per-subject embeddings were cached to disk so that repeat head-training runs incurred no redundant forward passes. Hyperparameter sweeps were budgeted ahead of time and pruned aggressively when fold-zero results indicated no signal; ensembling and large-batch search were deferred or skipped where the incremental compute did not justify the expected gain. No experiment in this dissertation required dedicated cluster time beyond what was already allocated to the research group.

## Availability of Data and Materials

All source code, configuration files, experiment manifests, and analysis notebooks for this dissertation are available in the project repository at https://github.com/abhivir-42/ai-for-fetal-heart-ultrasound. The iFIND fetal ultrasound corpus is governed by the iFIND research project's data-sharing agreement and is not redistributable; access requests should be directed to the iFIND principal investigators at Imperial College London. Per-subject embeddings, per-fold predictions, calibration outputs, and the figures and tables in this report are reproducible from the released code given the iFIND raw data and the documented environment specification.

# Appendix A

# Supplementary Material

This appendix collects supplementary material that supports the main report without being load-bearing for the central argument: most prominently, the full Collins 2024 TRIPOD+AI item-by-item compliance table (Section A.1). The appendix is not required reading; the main report is intended to make sense if the appendix is removed.

## A.1 TRIPOD+AI compliance table

This appendix audits the dissertation against the 27-item TRIPOD+AI 2024 checklist [10]. The numbering and item titles follow the canonical Collins et al. 2024 statement, which supersedes the 2015 TRIPOD checklist and is the relevant reporting standard for clinical prediction model studies that use machine learning methods. Sub-items (e.g. 3a, 5a, 12a) are listed where the canonical checklist itself enumerates them. The status column uses three codes: MET (the dissertation fully addresses the item), PARTIAL (the item is addressed but with explicit qualification or limitation), and UNMET (the item is not addressed; the principal unmet items are named in the summary paragraph below). The "where addressed" column points at the chapter, section, table, or declaration that addresses each item.

| # | Item (Collins 2024) | Status | Where addressed (chapter / section / table / declaration) |
|---|---|---|---|
| 1 | Title — identify the study as developing or evaluating a multivariable prediction model, specifying target population and outcome. | MET | Titlepage of `report.tex`: 'Fetal Heart Ultrasound Congenital Heart Disease Classification on the iFIND Dataset'; identifies prediction-model study, target population (second-trimester fetal ultrasound), and outcome (CHD presence). |
| 2 | Abstract — structured summary per TRIPOD+AI for Abstracts. | MET | Abstract (`report.tex`): study design, predictors, outcome, sample size (4,128 subjects), discrimination and calibration results, dual finding, key limitation (internal-only evaluation). |
| 3a | Background — healthcare context and rationale; references to existing models. | MET | Sections 1.2 (motivation) and 2.1 (clinical context: fetal cardiac screening); references to prior fetal-cardiology screening literature throughout Chapter 2. |

*Continued on next page.*

*Table A.1 continued from previous page.*

| # | Item (Collins 2024) | Status | Where addressed (chapter / section / table / declaration) |
|---|---|---|---|
| 3b | Background — target population, intended purpose within care pathway, intended users. | MET | Section 1.1 (project statement); Section 7.4 (deployment recipe, intended users: tertiary-referral screening gate); Section 2.1 (UK FASP care pathway). |
| 3c | Background — known health inequalities between sociodemographic groups. | PARTIAL | Section 2.1 acknowledges UK FASP detection-rate variation across centres and operator experience as a known inequality driver; sub-population disparities by maternal demographics are not enumerated and are flagged under Section 7.4 (fairness across demographic groups). |
| 4 | Objectives — specify study objectives (development / evaluation / both). | MET | Section 1.3 (research questions and aims); study is development plus internal evaluation. |
| 5a | Data — describe data sources separately for development and evaluation, with rationale and representativeness assessment. | MET | Section 3.1 (iFIND cohort, 4,128 subjects, single multi-site Imperial-affiliated programme); Section 2.2 (iFIND project provenance); subject-disjoint train / validation / test partitions of Section 3.6 furnish the development-versus-evaluation separation within a single cohort. |
| 5b | Data — dates of data collection, participant accrual, follow-up. | PARTIAL | Section 3.1 describes the iFIND programme accrual qualitatively; precise accrual-window dates and per-clip acquisition dates are not enumerated subject by subject because the released cohort is a fixed retrospective resource. |
| 6a | Participants — study setting, number and location of centres. | MET | Section 3.1 (Imperial-affiliated hospitals, retrospective collection during routine ∼20-week anomaly scans); Section 2.2 (iFIND multi-centre context). |
| 6b | Participants — eligibility criteria. | MET | Sections 3.1 and 3.6: inclusion is by second-trimester anomaly scan acquisition within the iFIND programme; subject-disjoint k-fold splitting with locked test partition. |
| 6c | Participants — treatments received and handling during model development or evaluation. | MET | Not applicable in the usual sense (the study is a diagnostic prediction study at the screening scan, not a treatment-conditional prognostic model); the absence of treatment-conditional design is made explicit in Section 1.1 (the prediction target is the antenatal scan moment, prior to any intervention). |
| 7 | Data preparation — preprocessing, quality checking, consistency across sociodemographic groups. | PARTIAL | Sections 4.3 (per-clip embedding pipeline), Section 5.2 (FetalCLIP backbone-swap specifics), Section 3.4 (postnatal-confirmation gating, CoA cleanup); data quality is checked at the cohort level but not stratified by sociodemographic group (flagged in Section 7.4). |

*Continued on next page.*

*Table A.1 continued from previous page.*

| # | Item (Collins 2024) | Status | Where addressed (chapter / section / table / declaration) |
|---|---|---|---|
| 8a | Outcome — definition, time horizon, assessment methods, consistency across sociodemographic groups. | MET | Sections 3.2 (9-condition labelling schema), 3.3 (auditable-4 lumped binary), 3.5 (per-condition slicing-$i$), and 3.8 (label-quality caveats); outcome is antenatal-suspicion CHD at the second-trimester scan moment. |
| 8b | Outcome — assessor qualifications and demographics where subjective interpretation is required. | PARTIAL | Section 3.2 records labels are derived from iFIND clinical team antenatal scan reports; Section 3.4 documents the postnatal-confirmation gating for CoA in consultation with the fetal cardiologist (Tom Day); individual sonographer-level qualifications and demographics are not reported subject-by-subject. |
| 8c | Outcome — blinding of outcome assessment. | PARTIAL | Outcome labels are taken from clinical scan reports recorded prospectively at the time of the original screening scan, before any model development; assessor blinding to model predictions is therefore satisfied by temporal order. Explicit blinding-of-outcome-assessment statements per CONSORT-style language are not provided. |
| 9a | Predictors — choice of initial predictors and any pre-selection. | MET | Sections 4.2–4.4 (DINOv2 Tier 1 architecture family); Section 5.2 (FetalCLIP Tier 2 backbone-swap); predictors are frozen-backbone embeddings of the ultrasound video, with no manual pre-selection. |
| 9b | Predictors — predictor definitions, measurement timing, blinding. | MET | Sections 4.3 (per-clip embedding pipeline, $nf{=}10 \times nc{=}20$ canonical token budget); predictors are extracted from the same antenatal scan that produced the outcome label, so prediction is conditioned on data available at the screening moment. |
| 9c | Predictors — subjective interpretation assessor qualifications. | MET | Not applicable: predictors are pixel-derived embeddings from frozen vision backbones, not subjective human interpretations. |
| 10 | Sample size — explain how study size was determined, separately for development and evaluation. | PARTIAL | Section 3.1 reports the 4,128-subject cohort and per-condition positive counts (Section 3.5); a formal Riley et al. four-component sample-size calculation [25] is not performed because the iFIND cohort is a fixed retrospective resource, and the post-hoc compensation is the discipline of reporting wide per-condition BCa CIs at $n_{\text{pos}} < 6$ (Section 5.5). Discussed honestly in Section 7.3. |
| 11 | Missing data — handling, reasons for omission. | MET | Section 3.4 (postnatal-confirmation gating via `tok[5]`, with cleaned-cohort vs uncleaned-cohort auditing in Section 5.6); per-subject missing-clip handling is described in Section 4.3. |

*Continued on next page.*

*Table A.1 continued from previous page.*

| # | Item (Collins 2024) | Status | Where addressed (chapter / section / table / declaration) |
|---|---|---|---|
| 12a | Analytical methods — data partitioning for development and evaluation. | MET | Section 3.6 (subject-disjoint train / validation / test splits with locked test partition); five-fold cross-validation discipline applied throughout Chapters 4–6. |
| 12b | Analytical methods — predictor handling, functional form, transformations. | MET | Sections 4.2–4.4 (Tier 1: frozen DINOv2 ViT-B/14 plus head architectures E2, F2a, F2b, B2, C1); Section 5.2 (Tier 2: FetalCLIP backbone-swap); no manual predictor transformations beyond mean-pool / attention-pool / cross-attention pooling per head. |
| 12c | Analytical methods — model type, rationale, building steps, hyperparameter tuning, internal validation. | MET | Sections 4.4–4.5 (head architectures and training recipe); Section 5.1 (Tier 2 evaluation setup); hyperparameter sweep discipline and the WS1 scaling investigation documented in Section 5.7. |
| 12d | Analytical methods — handling and quantification of heterogeneity across clusters. | PARTIAL | Per-fold reporting discipline (Sections 5.3 and 6.3) surfaces between-fold variance; centre-level or scanner-level heterogeneity is not analysed separately and is flagged in Section 7.3 as a partial confounder analysis. |
| 12e | Analytical methods — performance measures and comparison plots. | MET | AUROC with paired-bootstrap 5-fold CIs and BCa method [23] (Sections 4.5, 5.3); Brier–Murphy decomposition [22] and reliability diagrams (Section 6.4); per-fold decision-curve analysis [3] at $p_t \in \{0.10, 0.15, 0.20\}$ (Section 6.3); temperature scaling [5] (Section 6.5). |
| 12f | Analytical methods — model updating or recalibration. | MET | Section 6.5 (cross-fold temperature scaling as the recalibration intervention, with full per-fold tables); the T-scaling intervention is reported as a calibrated-probability mechanism that halves but does not close the per-fold gap on auditable-4. |
| 12g | Analytical methods — how predictions were calculated for evaluation. | MET | Sections 4.4–4.5 (subject-level prediction via mean-pool / attention-pool over the per-clip embedding sequence, followed by head-specific aggregation and sigmoid); Section 5.2 (Tier 2 prediction pipeline). |
| 13 | Class imbalance — if used, state why and how, plus recalibration. | MET | Section 3.2 reports the 15.50% any-CHD prevalence and the ∼10% auditable-4 prevalence; no resampling or class-weighting is used during head training (binary-cross-entropy loss on the natural prevalence); post-training recalibration via temperature scaling is reported in Section 6.5. |

*Continued on next page.*

*Table A.1 continued from previous page.*

| # | Item (Collins 2024) | Status | Where addressed (chapter / section / table / declaration) |
|---|---|---|---|
| 14 | Fairness — approaches to address model fairness and rationale. | PARTIAL | Section 7.4 (responsible-AI framing) acknowledges that sub-population disparities by maternal demographics are not measured; subject-disjoint k-fold splitting is reported but demographic-stratified evaluation is not performed. Section 7.3 (threats to validity) flags the same limitation. |
| 15 | Model output — prediction output type and rationale for classification and threshold identification. | MET | Sections 1.3, 4.5, and 6.2: output is a calibrated probability of any-CHD (or auditable-4 CHD); the screening operating point is $p_t \in \{0.10, 0.15, 0.20\}$ as the pre-registered DCA grid (rationale: UK FASP-aligned screening context, Section 2.1). |
| 16 | Training versus evaluation — identify differences between development and evaluation data in setting, eligibility, outcome, predictors. | PARTIAL | Section 3.6 reports the subject-disjoint internal split; no external-cohort evaluation is performed, so training-versus-evaluation differences in setting (other centres, other scanners, other sonographer populations) are not characterised. Flagged as the principal future-work item in Section 7.6 (PROMETHEUS external validation). |
| 17 | Ethical approval — name the institutional ethics committee and describe consent or waiver. | MET | Declarations: `ethics.tex`; Section 1.5 (ethical considerations); Section 7.4 (data governance). Retrospective de-identified-data study under the existing iFIND institutional ethics approval; no new human-participant intervention. |
| 18a | Open science — funding source and funder roles. | MET | Declarations: `report.tex` (Sustainability section); compute resources from Imperial Department of Computing and a collaborator-provided four-GPU workstation; no external project funding declared. |
| 18b | Open science — conflicts of interest and financial disclosures. | MET | Declarations: `ethics.tex` discloses the supervisor's co-investigator role on iFIND; `genai-declaration.tex` discloses generative-AI tool use; no other financial conflicts declared. |
| 18c | Open science — study protocol accessibility. | PARTIAL | No formally registered prospective protocol exists (the study is a retrospective computational dissertation on a fixed cohort); the methodological discipline, pre-registered DCA grid, and per-fold replication criteria are documented inline in Chapters 4–6 and in the source-of-truth documentation referenced from the project repository. |
| 18d | Open science — registration. | UNMET | The dissertation is not pre-registered on a prospective prediction-model registry; no equivalent register is conventional for retrospective computational dissertations of this scope. |

*Continued on next page.*

*Table A.1 continued from previous page.*

| # | Item (Collins 2024) | Status | Where addressed (chapter / section / table / declaration) |
|---|---|---|---|
| 18e | Open science — data availability. | MET | Declarations: `report.tex` (Availability of Data and Materials); the iFIND raw cohort is governed by the iFIND data-sharing agreement and is not redistributable; per-subject embeddings, per-fold predictions, calibration outputs, and the figures and tables in the report are reproducible from the released code given the iFIND raw data and the documented environment. |
| 18f | Open science — analytical code availability. | MET | Declarations: `report.tex` (Availability of Data and Materials); all source code, configuration files, experiment manifests, and analysis notebooks are at https://github.com/abhivir-42/ai-for-fetal-heart-ultrasound. |
| 19 | Patient and public involvement — design, conduct, reporting, interpretation, dissemination. | PARTIAL | The Declarations (Use of Generative AI; Ethical Considerations) record consultation with the project's clinical authority, Dr Tom Day, on clinical framing, the auditable-four selection, and the screening operating-point grid; no formal patient-and-public-involvement panel was convened for the dissertation. |
| 20a | Participants — flow through study, outcome event numbers, follow-up summary. | MET | Section 3.1 (4,128 → 4,086 subjects after CoA cleanup); Section 3.5 (Table 3.3, per-condition positive counts); Section 3.6 (k-fold partition flow). |
| 20b | Participants — characteristics overall and by data source. | MET | Section 3.2 (Table 3.1); Section 3.4 (Table 3.2); Section 3.5 (Table 3.3). |
| 20c | Participants — for evaluation studies, compare development data distribution of important predictors. | PARTIAL | Internal-only evaluation: distributions of cohort properties are reported for the full cohort but no cross-cohort predictor distribution comparison is performed (no external cohort evaluated). Flagged in Section 7.6. |
| 21 | Model development — participant and outcome event numbers in each analysis (development, hyperparameter tuning, evaluation). | MET | Section 3.6 (split sizes); per-fold positive-count tables in Sections 4.5–4.7 and 5.3; WS1 scaling-investigation positive counts in Section 5.7. |

*Continued on next page.*

*Table A.1 continued from previous page.*

| # | Item (Collins 2024) | Status | Where addressed (chapter / section / table / declaration) |
|---|---|---|---|
| 22 | Model specification — full prediction model details enabling new predictions and third-party evaluation; access restrictions. | MET | Section 4.4 (head architecture family E2, F2a, F2b, B2, C1, with full parameter counts and architectural diagrams in Chapter 4); Section 5.2 (FetalCLIP backbone-swap); released code, configuration files, and trained weights at the project repository (`Availability of Data and Materials` declaration); re-prediction on the iFIND cohort is fully reproducible. |
| 23a | Model performance — estimates with confidence intervals, including key subgroup results. | MET | Sections 4.5, 4.7 (Tier 1 all-9 and auditable-4 AUROC with paired-bootstrap 5-fold BCa CIs); Sections 5.3–5.5 (Tier 2 backbone-swap result plus per-condition slicing-$i$); Section 6.3 (per-fold decision-curve analysis at the pre-registered screening grid); Section 6.4 (Brier–Murphy decomposition with reliability term). |
| 23b | Model performance — heterogeneity across clusters. | PARTIAL | Per-fold variance is reported throughout Chapters 4–6 (per-fold NB discipline, per-fold AUROC); centre-level or scanner-level heterogeneity is not analysed separately because the iFIND cohort is a single multi-site Imperial-affiliated programme and per-centre metadata are not used. Flagged in Section 7.3. |
| 24 | Model updating — results from any model updating including updated model and subsequent performance. | MET | Section 6.5 (temperature-scaling intervention as the recalibration update; full per-fold table at `experiments/2026-05-26-tscale-verification/results.md`); Section 7.1 (the auditable-4 T-scaling qualification and its honest accounting). |
| 25 | Interpretation — main results, fairness considerations, context. | MET | Section 7.1 (synthesising the dual finding); Section 7.1 (threshold-vs-aggregate calibration generalisation); Section 7.4 (fairness considerations); Chapter 8 (closing interpretation). |
| 26 | Limitations — study limitations and their effects on bias, statistical uncertainty, and generalisability. | MET | Section 7.3 (threats to validity: antenatal-suspicion labelling, per-condition power, single-cohort selection, per-fold variance, WS1 scaling null at the wrong scale, operating-point sensitivity, partial confounder analysis); Section 3.8 (label-quality caveats). |
| 27a | Usability and next steps — assessment and handling of poor quality or unavailable input data. | PARTIAL | Section 4.3 (per-clip embedding pipeline handles variable per-subject clip counts); a formal at-inference quality-check pipeline is not specified because the dissertation does not deploy a clinical system. Flagged under Section 7.4 (deployment recipe) as a deployment-side requirement. |

*Continued on next page.*

*Table A.1 continued from previous page.*

| # | Item (Collins 2024) | Status | Where addressed (chapter / section / table / declaration) |
|---|---|---|---|
| 27b | Usability and next steps — required user interaction and expertise for implementation. | PARTIAL | Section 7.4 sketches the two-head deployment recipe (DINOv2 mean-pool linear probe as screening gate, post-T-scaled FetalCLIP F2a as calibrated probability estimator) and identifies tertiary-referral screening as the intended use; full human-factors and user-interaction specifications are out of scope for a dissertation. |
| 27c | Usability and next steps — future research on applicability and generalisability. | MET | Section 7.6 sets out the future-work directions — the asymptotic-budget scaling run, multi-window motion sampling, a clip-level pretext task, condition-specific heads, a demographic-stratified fairness evaluation, and external validation on PROMETHEUS — as the highest-leverage next steps. |

**Status summary.** Of the 52 rows above (the 27 parent items of the canonical TRIPOD+AI checklist, with sub-items enumerated wherever the published statement itself sub-divides them), 36 are MET, 15 are PARTIAL, and 1 is UNMET. The single UNMET item is Item 18d (registration on a prospective prediction-model registry), which is not conventional for a retrospective computational dissertation of this scope. The PARTIAL items cluster around five substantive gaps: (i) the absence of an **external-cohort evaluation** (Items 5b, 16, 20c), which is the dissertation's single highest-leverage future-work item and the gating prerequisite for any clinical-deployment claim (Section 7.6, PROMETHEUS pathway); (ii) the absence of a **formal Riley et al. sample-size calculation** (Item 10), with the post-hoc compensation of wide per-condition BCa CIs at $n_{\text{pos}} < 6$ (Section 5.5); (iii) the absence of **sub-population fairness analysis by maternal demographics** (Items 3c, 7, 14, 23b), acknowledged as a limitation in Section 7.4; (iv) the absence of **formal PPI and protocol registration** (Items 18c, 19), with the clinical-collaborator consultation (Tom Day, fetal cardiologist) reported instead; and (v) **outcome-assessment formalities and deployment-side usability** (Items 8b, 8c, 12d, 27a, 27b), where the substantive content is reported but CONSORT-style blinding statements and a deployed clinical user-interaction specification are out of scope for the dissertation. All **load-bearing methodological items** (Items 5a, 6a–c, 8a, 9a–c, 11, 12a–c, 12e–g, 13, 15, 17, 18a, 18b, 18e, 18f, 20a, 20b, 21, 22, 23a, 24, 25, 26, 27c) are MET, with the discrimination assessment (paired-bootstrap BCa CIs), calibration assessment (Brier–Murphy decomposition plus temperature scaling), and clinical-utility assessment (per-fold decision-curve analysis at the pre-registered screening grid) all reported per the TRIPOD+AI evidence standard.

**Note on numbering.** The item numbering above follows the canonical Collins et al. 2024 TRIPOD+AI statement [10], in which Item 5 is *Data*, Item 6 is *Participants*, Item 10 is *Sample size*, Item 22 is *Model specification*, Item 26 is *Limitations*, and external validation is addressed jointly under Items 16, 20c, 23b, 26, and 27c. Sections 2.7 and 7.5 of the main report use this canonical Collins 2024 numbering consistently; this table is the authoritative cross-reference index.