Delft University of Technology

A DIAGNOSTIC FRAMEWORK FOR ANNOTATION SHIFT IN CROSS-DOMAIN MACHINE LEARNING

Ralitsa Ianatchkova

Student id: 6159710

Pattern Recognition and Bioinformatics Research Group
Thesis Advisor: Jesse Krijthe
Daily Supervisor: Gijs van Tulder
External Committee Member: Emir Demirović

Delft
22nd of June, 2026

A Diagnostic Framework for Annotation Shift in Cross-Domain Machine Learning

Ralitsa Ianatchkova

Pattern Recognition and Bioinformatics Research Group, TU Delft

Abstract

This paper introduces a diagnostic framework for assessing annotation shift in cross-domain machine learning, with a focus on medical imaging applications. We formally define annotation shift as a change in the conditional distribution of assigned labels given the underlying target state. This distinction separates annotation-related effects from prevalence and acquisition-related shifts, which may produce similar observable patterns. We develop a framework combining input-distribution diagnostics, label-distribution analysis, and bidirectional cross-domain model evaluation to assess whether observed differences are consistent with annotation shift. The approach is evaluated through controlled synthetic experiments and experiments using osteoarthritis radiographs. Across both settings, annotation shift produces characteristic directional asymmetries in cross-domain prediction errors that differ from the signatures of prevalence and acquisition shifts. These asymmetries provide a basis for distinguishing annotation shift from other forms of domain shift, enabling more reliable interpretation of cross-domain model failures.

1 Introduction

Machine learning models for medical imaging often perform well in controlled settings, but degrade when deployed across domains, such as different institutions, cohorts, or data-generating environments [1, 2]. This degradation is typically caused by dataset shift: differences between the development and deployment domains that affect the relationship between inputs, labels, and underlying clinical states [3]. The source of this shift may lie, for instance, in the input data, differences in disease prevalence or severity, or differences in the labelling process [4]. These mechanisms call for different responses. For example, input differences may require harmonisation, prevalence differences may require recalibration or reweighting, and labelling differences may require changes to annotation protocols or label interpretation [5].

This work focuses on annotation shift, the possibility that the labelling process itself differs across domains. Annotation shift can arise whenever labels are assigned through human judgement, institutional conventions, or imperfect measurement rather than directly observed ground truth. Differences in grading conventions, decision thresholds, annotator expertise, or labelling protocols may induce systematic differences in how the same underlying case is labelled [6]. This issue is especially important in medical applications, where labels often

reflect subjective expert interpretation and may therefore vary between institutions, clinical settings, or annotators. For example, two hospitals may apply different thresholds for assigning a borderline imaging case to a more severe category, even when the underlying clinical state is the same.

To study annotation shift, it is common to begin from the observed relationship between inputs and labels. In dataset-shift terminology, changes in the conditional distribution $P(Y \mid X)$ across domains are often described as concept shift, where X denotes the observed input, such as a medical image, and Y denotes the assigned label [7]. Because X and Y are typically the quantities available in observed data, changes in $P(Y \mid X)$ provide a practical way to study whether labelling behaviour may vary across domains. However, such changes do not uniquely identify annotation shift, since they may also arise from differences in the input distribution, disease prevalence or severity, or the way the same underlying disease state is represented in the observed image [8].

This ambiguity motivates distinguishing between the observed assigned label and the underlying target state that the label is intended to represent. In most datasets, the assigned label is only an imperfect measurement or interpretation of an underlying state, such as disease severity [9]. We therefore distinguish between Y_{label}, the observed assigned label, and Y_{true}, the underlying target state, and define annotation shift as a change in $P(Y_{\text{label}} \mid Y_{\text{true}})$. This definition captures the idea that annotation shift concerns changes in how the same underlying state is labelled, while separating labelling behaviour from changes in prevalence or input representation. The theoretical implications of this distinction are developed in Section 3.

Since Y_{true} is typically unobserved, annotation shift cannot usually be evaluated directly. In practice, we must rely on observable quantities: the input X and the assigned label Y_{label}. In this work, we develop a diagnostic framework that combines input-distribution diagnostics, label-distribution diagnostics, and bidirectional cross-domain model evaluation to assess whether observed cross-domain differences are more consistent with annotation shift or with alternative explanations such as acquisition-related variation or prevalence differences. Rather than attempting to prove the presence of annotation shift, the goal is to determine when annotation shift is a plausible explanation for domain shift.

We develop this diagnostic framework in the context of hip osteoarthritis severity grading from X-ray images using the Kellgren–Lawrence system, an ordinal grading scheme known to exhibit inter-rater variability particularly at lower severity levels [10], though the diagnostic perspective applies more broadly to subjective labelling tasks.

2 Related Work

2.1 Dataset Shift

Dataset shift is commonly divided into three main categories: covariate shift, where $P(X)$ changes; prior-probability shift, where $P(Y)$ changes; and concept

shift, where the relationship between inputs and labels $P(Y \mid X)$ differs across domains [7, 8].

Castro et al. [5] extend this framing to medical imaging from a causal perspective, distinguishing prevalence shift, where the distribution of the underlying disease state or case mix changes, from acquisition shift, where the imaging process or visual appearance of the data changes. In the notation used here, these correspond to changes in Y_{true} and changes in the representation of Y_{true} in the observed input X, respectively.

Castro et al. also introduce annotation shift to describe differences in how the same underlying case is labelled across domains. Their formulation captures the important idea that labelling mechanisms may differ across institutions, annotators, or annotation protocols. However, when annotation shift is expressed as a change in $P(Y \mid X)$, it overlaps with classical concept shift, since both describe changes in the observed relationship between inputs and labels.

Moreover, although Castro et al. conceptually distinguish mechanisms involving the underlying disease state from mechanisms involving the assigned label, these two roles of Y are not formally separated in the notation. In this work, we make this distinction explicit by separating the latent target state Y_{true} from the observed label Y_{label}, and define annotation shift as a change in $P(Y_{\text{label}} \mid Y_{\text{true}})$. This isolates changes in labelling behaviour from changes in disease prevalence, image acquisition, or disease representation.

2.2 Shift Detection Methods

Prior work addresses the detection of dataset shift through two main methodological paradigms: distribution-based methods, which compare the statistical distributions of observed data across domains, and model-behaviour-based methods, which examine how predictive models trained in one domain perform or make errors in another [11].

2.2.1 Distribution-based Methods

A family of methods detects dataset shift by directly comparing source and target distributions. Some approaches use statistical discrepancy measures to compare distributions directly. For example, Gretton et al. [12] propose a kernel two-sample test based on Maximum Mean Discrepancy (MMD), which tests whether two samples are likely to have been drawn from the same distribution by comparing their kernel mean embeddings. Other approaches formulate shift detection as a classification problem, which is particularly common in high-dimensional settings such as medical imaging. Lopez-Paz and Oquab [13] propose a classifier two-sample test, where a classifier is trained to predict whether an example comes from the source or target domain; above-chance performance indicates that the domains are statistically distinguishable. Related work also considers settings where the class proportions change between domains while $P(X \mid Y)$ is assumed to remain stable. For example, Lipton et al. [14] show

that target-domain class proportions can be estimated from black-box classifier predictions even without target labels.

Several frameworks combine multiple distribution-based tests. For example, DetectShift [15] provides a unified framework for testing shifts in several observable distributions, including $P(X)$, $P(Y)$, $P(X \mid Y)$, $P(Y \mid X)$, and the joint distribution $P(X, Y)$. In the terminology used above, a shift in $P(X)$ corresponds to covariate shift, a shift in $P(Y)$ corresponds to prior-probability shift, and a shift in $P(Y \mid X)$ corresponds to concept shift. Thus, such methods assess dataset shift through statistical comparisons of observable distributions, rather than through downstream model behaviour. In the medical-imaging setting, a detected shift in $P(X)$ may suggest acquisition or preprocessing differences, while a shift in $P(Y)$ may reflect differences in observed label proportions.

Under the definition proposed in this work, however, distribution-based methods remain limited by their reliance on observable variables. They can detect changes in $P(X)$, $P(Y_{\text{label}})$, or $P(Y_{\text{label}} \mid X)$, but annotation shift is defined as a change in $P(Y_{\text{label}} \mid Y_{\text{true}})$. Because Y_{true} is unobserved, a detected change in $P(Y_{\text{label}} \mid X)$ does not uniquely identify annotation shift. Instead, it may also arise from acquisition-related variation, differences in disease prevalence, or differences in how the same underlying disease state is represented in the observed input. In this work, we therefore use distribution-based methods primarily to assess observable input differences and to evaluate alternative explanations for the patterns observed in model behaviour.

2.2.2 Model-behaviour-based Methods

Another way to study dataset shift is through cross-domain model evaluation. A common approach is to train a model in one domain and evaluate it on data from another, comparing cross-domain performance with within-domain performance. For example, Zech et al. [16] evaluate pneumonia-detection models across hospital systems to assess how performance changes when the deployment domain differs from the training domain. Cohen et al. [17] and Pooch et al. [18] use related multi-dataset evaluation strategies for chest X-ray prediction, comparing how models trained on one or more datasets generalise to held-out external datasets.

Under the definition proposed in this work, however, these methods remain indirect. They operate on observable quantities, namely X and Y_{label}, and therefore provide evidence about changes in $P(Y_{\text{label}} \mid X)$. Under our definition, annotation shift is a change in $P(Y_{\text{label}} \mid Y_{\text{true}})$, which cannot be observed directly because Y_{true} is unavailable. Consequently, cross-domain performance degradation or systematic error patterns may be consistent with annotation shift, but they do not distinguish annotation-related differences from acquisition-related or prevalence-related differences. In this work, we therefore combine model-behaviour diagnostics with input-distribution comparisons. Specifically, we use cross-domain performance and directional error patterns as observable signatures, while separately assessing whether detectable differences in the input distribution provide an alternative explanation for the observed model behaviour.

2.3 Adapting to Dataset Shift

Methods for adapting to dataset shift typically depend on assumptions about which component of the data-generating process has changed. Approaches targeting covariate or acquisition shift often aim to learn domain-invariant representations or align source and target feature distributions [19]. Under shifts in class proportions, adaptation methods instead commonly rely on estimating target-domain label distributions and reweighting predictions or training samples accordingly [20]. When differences arise from the labelling process itself, adaptation may require revised annotation protocols, consensus procedures, or explicit modelling of annotator behaviour [21]. These differences are important because an adaptation strategy that is appropriate for one shift mechanism may be ineffective or misleading under another. This motivates the need for diagnostic approaches that can distinguish between competing explanations for observed cross-domain differences.

3 Theoretical Analysis of Annotation Shift

The central idea behind annotation shift is that the same underlying case may receive different observed labels depending on the annotator, institution, cohort, or labelling protocol. This suggests that annotation shift should be understood as a change in the mapping from an underlying target state to an observed label, rather than only as a change in the observed relationship between inputs and labels.

3.1 Formal Definition of Annotation Shift

Let X denote the observed input, Y_{label} the label assigned by an annotator or labelling process, and Y_{true} the underlying target state that the label is intended to represent. The underlying target state need not have the same form as the observed label: it may be continuous, ordinal, or categorical, while Y_{label} may be a discretised or otherwise imperfect measurement of it. In medical settings, Y_{true} may correspond to the true disease state, disease severity, or clinical outcome of interest.

> **Definition.** We define annotation shift as a change in $P(Y_{\text{label}} \mid Y_{\text{true}})$ across domains.

Under this definition, annotation shift concerns differences in how the same underlying target state is translated into an observed label, as illustrated in Figure 1. For example, a more experienced radiologist may grade more accurately relative to this target state, while a policy change may redefine which disease threshold triggers a positive label. Neither case necessarily involves a change in the observed input X or in the underlying target state Y_{true}, but both would result in a change in the observed label Y_{label}.

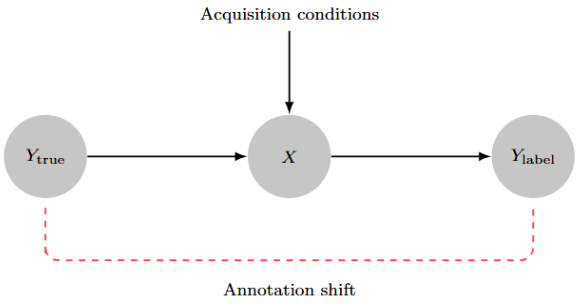


Figure 1: Schematic illustration of annotation shift, where differences arise in the mapping from the latent target state Y_{true} to the observed label Y_{label}, while acquisition conditions and the observed input representation X remain stable.

This formulation separates the labelling process from other sources of cross-domain variation. In some applications, Y_{true} may be approximated by a more reliable reference standard, such as biopsy results, longitudinal follow-up, or expert adjudication. In many settings, however, such a reference is unavailable, and annotation shift must be studied indirectly through observable variables.

3.2 Why $P(Y_{\text{label}} \mid X)$ is Insufficient

While prior work often characterises annotation shift through changes in $P(Y_{\text{label}} \mid X)$, the definition proposed here instead focuses on changes in $P(Y_{\text{label}} \mid Y_{\text{true}})$. The question then becomes under what conditions observable changes in $P(Y_{\text{label}} \mid X)$ can still be interpreted as evidence for annotation shift. A change in $P(Y_{\text{label}} \mid X)$ is interpretable as evidence for annotation shift only under additional assumptions about what remains stable across domains. In particular, the observed input must be a comparable representation of the underlying target state. Formally, this requires an assumption such as

$$P_A(X \mid Y_{\text{true}}) = P_B(X \mid Y_{\text{true}}) \tag{1}$$

meaning that the same underlying target state gives rise to comparable observed inputs in domains A and B.

A concrete example is radiologist adaptation to scanner characteristics, as depicted in Figure 2. When presented with images from a different institution, a radiologist's labelling strategy may appear to shift not because their grading policy has changed, but because the visual presentation of the same pathology differs. This would change $P(Y_{\text{label}} \mid X)$ through a change in $P(X \mid Y_{\text{true}})$, and methods that test for changes in the observed label-input relationship could therefore interpret it as a shift, even though the mapping from Y_{true} to Y_{label} has not changed. As a result, an analysis based solely on $P(Y_{\text{label}} \mid X)$ would incorrectly suggest the presence of annotation shift.

Conversely, if two domains contain comparable images for the same underlying disease states, but one institution systematically assigns more severe grades

to borderline cases, a change in $P(Y_{\text{label}} \mid X)$ is more plausibly explained by annotation shift. The distinction therefore depends not only on whether the label-input relationship changes, but also on whether the input remains comparable as a representation of the underlying target state.

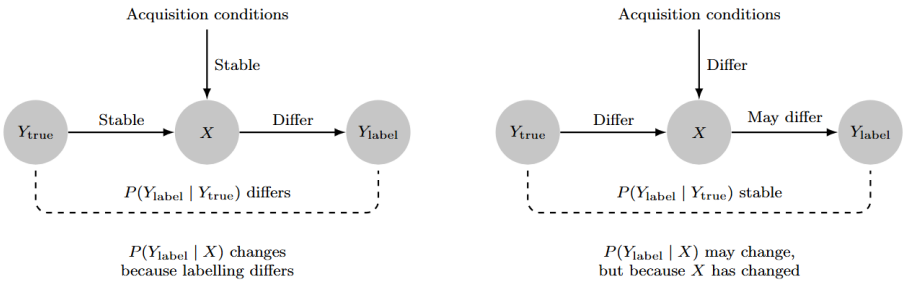


Figure 2: Illustration of how changes in the observed relationship $P(Y_{\text{label}} \mid X)$ may arise from different mechanisms. Left: annotation shift, where $P(Y_{\text{label}} \mid Y_{\text{true}})$ differs while acquisition conditions remain stable. Right: acquisition-related shift, where $P(Y_{\text{label}} \mid X)$ may change because the representation of Y_{true} in X changes, while $P(Y_{\text{label}} \mid Y_{\text{true}})$ remains stable.

3.3 Using Observable Quantities in Practice

If Y_{true} is unobserved, annotation shift cannot be measured directly and must instead be studied through observable quantities. This raises the question of how changes in $P(Y_{\text{label}} \mid X)$ should be interpreted in practice.

A first consideration is whether the observed input distribution appears stable across domains. If no detectable difference in $P(X)$ is found, this reduces support for acquisition-related explanations of a change in $P(Y_{\text{label}} \mid X)$. In that case, observed differences in the label-input relationship are more plausibly attributed to differences in prevalence or labelling behaviour.

If differences in $P(X)$ are detected, one approach is to reduce acquisition-related variation through harmonisation or a transformation $f(X)$ such that

$$P_A(f(X) \mid Y_{\text{true}}) \approx P_B(f(X) \mid Y_{\text{true}}) \tag{2}$$

For example, f may correspond to image normalisation, scanner harmonisation, feature standardisation, or a learned representation designed to reduce institution-specific acquisition effects. If such a transformation makes comparable underlying target states appear more similar across domains, then differences in $P(Y_{\text{label}} \mid f(X))$ provide stronger evidence for annotation shift than differences observed in the original input space.

4 Methodology

Rather than attempting to prove the presence of annotation shift directly, the goal of this framework is to assess whether observed cross-domain differences are more consistent with annotation shift or with alternative explanations such as acquisition-related variation or prevalence differences. This requires evaluating whether the observed inputs provide sufficiently comparable representations of the underlying target state for changes in the observed label-input relationship to plausibly reflect changes in the labelling process itself. Rather than relying on a single metric, the framework combines observable diagnostics to assess which theoretical explanation is most consistent with the observed data.

4.1 Observable Diagnostics and Metrics

We consider two domains, denoted A and B. For each domain, we observe inputs X and assigned labels Y_{label}. The underlying target state Y_{true} is not observed. Cross-domain differences can be studied through three complementary sources of diagnostic evidence:

1. **Input-distribution diagnostics**, which test whether the domains differ in the observed inputs X.
2. **Label-distribution diagnostics**, which test whether the observed labels Y_{label} differ across domains.
3. **Bidirectional model-behaviour diagnostics**, which test how models trained in one domain behave when evaluated in the other domain.

4.1.1 Input-distribution Diagnostics

The input-distribution diagnostic assesses whether the two domains differ in their observed inputs X. We implement this using a classifier two-sample test [13], in which a domain classifier is trained to predict whether a sample originates from domain A or domain B. If the two domains are drawn from the same input distribution, the classifier should not generalise above chance on held-out data. Near-chance performance therefore suggests that there is no detectable difference in $P(X)$ for the classifier and representation used, while above-chance performance indicates an observable input-domain signal.

No domain signal. If no input-domain signal is detected, this suggests that the observed inputs are comparable for the classifier and representation used, but it does not prove that $P(X)$ is identical across domains. A domain classifier may fail to detect subtle input differences, and different shift mechanisms may also offset each other in the marginal distribution $P(X)$. Nevertheless, the absence of a detectable input-domain signal reduces support for acquisition-related explanations of subsequent cross-domain differences. In that case, model-behaviour diagnostics can be interpreted with greater focus on prevalence and annotation-related mechanisms.

Detected domain signal. Detectable differences in $P(X)$ may arise from acquisition-related variation, prevalence differences, or disease-relevant changes in how the underlying target state is represented. Annotation shift should not cause input differences. This matters because the same model-behaviour pattern that appears consistent with annotation shift may also be caused by differences in the observed inputs.

Persistent or non-removable domain signal. If an input-domain signal is detected, the next question is whether it can be reduced without removing information relevant to Y_{true}. Conceptually, this corresponds to a transformation $f(X)$ that suppresses domain-specific variation while preserving target-relevant information as discussed in Section 3.3. If such a transformation is available, the diagnostics should be repeated after correction. If the domain signal persists, or can only be removed at the cost of degrading target-relevant information, then the input-distribution diagnostic alone cannot rule out alternative explanations such as prevalence differences or disease-relevant acquisition shift. In that case, annotation shift must be interpreted using additional evidence, including label-distribution checks, bidirectional model behaviour, or external information.

4.1.2 Label-distribution Diagnostics

The label-distribution diagnostic assesses whether the observed label distributions $P(Y_{\text{label}})$ differ across domains. Such differences are informative because they may reflect either differences in the underlying disease prevalence or severity, or differences in labelling behaviour such as systematic over- or under-grading. A difference in $P(Y_{\text{label}})$ therefore indicates that the observed labelled datasets differ in some way, but it does not identify the source of this difference. Conversely, similar label distributions do not rule out prevalence or annotation shift, since changes in the underlying case mix and changes in labelling behaviour may offset each other.

4.1.3 Bidirectional Model-behaviour Diagnostics

The bidirectional model-behaviour diagnostic assesses how models trained in one domain behave when evaluated within and across domains. We train models separately in domains A and B, and evaluate them in four directions:

$$A \to A, \quad A \to B, \quad B \to B, \quad B \to A.$$

The within-domain evaluations, $A \to A$ and $B \to B$, provide reference performance under each domain's own data and labelling process. The cross-domain evaluations, $A \to B$ and $B \to A$, measure how well models transfer between domains.

This bidirectional design is important because systematic annotation differences, such as shifts in grading thresholds, may produce asymmetric cross-domain behaviour. For example, if domain B uses stricter labelling thresholds than domain A, a model trained on A may tend to assign higher labels than those observed in B, while a model trained on B may tend to assign lower

labels than those observed in A. Such a reversal is more informative than a one-directional performance drop, since ordinary domain shift can also reduce cross-domain performance. However, this signature is expected mainly for directional or threshold-like annotation shifts. Other changes in $P(Y_{\text{label}} \mid Y_{\text{true}})$, such as changes in annotator reliability, class-dependent noise, or non-monotone relabelling, may affect performance without producing a consistent upward or downward error direction.

Model behaviour is summarised using two complementary types of metrics.

First, **predictive performance** measures whether the model preserves useful task-relevant information across domains. The specific metric depends on the task setting: in binary experiments, this may be measured using accuracy, balanced accuracy, or AUROC, while in ordinal settings it may be measured using agreement or ordinal classification performance. Performance metrics are useful because they indicate whether the model continues to separate or classify cases across domains, but they do not by themselves identify the source of any degradation.

Second, we use **directional error rates** which measure whether cross-domain errors are systematically upward or downward in the ordinal scale. These are the most directly relevant behavioural signature for annotation shift.

For tasks with an ordered label space, including binary, ordinal, or continuous outcomes, prediction errors can be classified by their direction relative to the observed target-domain label. Let $\hat{Y}$ denote the predicted label or score and Y_{label} the observed label. An upward error occurs when $\hat{Y} > Y_{\text{label}}$, and a downward error occurs when $\hat{Y} < Y_{\text{label}}$. We summarise the directionality of errors using

$$\text{net direction} = P(\hat{Y} > Y_{\text{label}}) - P(\hat{Y} < Y_{\text{label}}) \tag{3}$$

A positive value indicates that the model tends to assign higher labels than the target-domain labels, while a negative value indicates that it tends to assign lower labels.

5 Diagnostic Framework

The diagnostics defined above can be combined to interpret cross-domain differences. The framework distinguishes annotation shift from alternative explanations such as acquisition-related variation or prevalence differences, since these mechanisms may produce similar degradations in cross-domain performance while requiring different interpretations and responses. Table 1 summarises the expected theoretical and empirical signatures of the main shift cases considered in the framework.

Table 1: Theoretical shift cases and their expected diagnostic signatures.

Case	**Theoretical change**	**Expected empirical pattern**	**Interpretation**
No shift	All components are stable	Domain-classifier performance near chance; similar label proportions; cross-domain predictive performance comparable to within-domain baselines; net direction small and balanced.	Reference condition.
Annotation shift only	$P(Y_{\text{label}} \mid Y_{\text{true}})$ and $P(Y_{\text{label}})$ change	Domain-classifier performance near chance; label proportions may differ; cross-domain predictive performance decreases relative to within-domain baselines; net direction reverses between transfer directions.	The strongest behavioural signature is a reversal in net direction: an A-trained model may over-grade on B, while a B-trained model under-grades on A, or vice versa.
Prevalence shift only	$P(Y_{\text{true}})$, $P(Y_{\text{label}})$ and $P(X)$ change	Domain-classifier performance may be above chance; label proportions differ across domains; cross-domain predictive performance may remain relatively preserved; net direction weak, symmetric, or inconsistent.	Suggests different case mix rather than different labelling conventions.
Prevalence and annotation shift	$P(Y_{\text{true}})$, $P(X)$ and $P(Y_{\text{label}} \mid Y_{\text{true}})$ change. $P(Y_{\text{label}})$ can be difficult to interpret because prevalence and annotation effects may reinforce or offset one another.	Domain-classifier performance may be above chance; label proportions may differ but are difficult to interpret; cross-domain predictive performance may decrease; net direction may reveal annotation differences but can be weakened or masked by prevalence effects.	Annotation shift could be present, but evidence may be confounded by prevalence differences.
Recoverable acquisition shift	$P(X \mid Y_{\text{true}})$ changes, but the change can be corrected by applying a harmonisation $f(X)$ without removing target-relevant information	Domain-classifier performance above chance before correction f; domain signal decreases after we apply f; label proportions remain similar; cross-domain predictive performance improves after correction; net direction weak or inconsistent.	Supports acquisition shift rather than annotation shift, unless annotation-like directional errors remain after correction.
Non-recoverable disease-relevant acquisition shift	$P(X \mid Y_{\text{true}})$ changes in a way that cannot be safely corrected while preserving target-relevant information.	Persistent domain-classifier signal in X; cross-domain predictive performance may degrade; net direction may resemble annotation shift even if labelling remains stable.	Annotation shift is not identifiable from observed data alone without external evidence or stronger assumptions.

The framework treats annotation-shift detection as a problem of comparing competing explanations for observed cross-domain differences. Rather than serving as definitive identification rules, the cases in Table 1 are intended as reference patterns for interpreting the joint diagnostic evidence. The combination of input-distribution diagnostics, label-distribution diagnostics, and bidirectional model-behaviour diagnostics is then used to determine which shift mechanism is most consistent with the observed data.

The table does not enumerate all possible combinations of shift mechanisms. In particular, combinations involving recoverable acquisition shift are treated through the correction step: if a valid transformation $f(X)$ can reduce acquisition-related variation while preserving target-relevant information, the diagnostics are repeated after applying $f(X)$, and any remaining patterns are interpreted using the non-acquisition cases. In practice, this requires checking that the domain-classifier signal decreases after applying $f(X)$, while within-domain task performance or other measures of target-relevant information remain comparable. By contrast, if the acquisition shift is non-recoverable and affects the disease-relevant representation, then annotation or prevalence effects cannot be reliably separated from input-related variation using observational data alone. For this reason, mixed cases involving non-recoverable acquisition shift are not treated as separately identifiable cases in the framework.

Figure 3 operationalises these cases as a sequential decision procedure, guiding the interpretation of observed diagnostics toward the most plausible shift mechanism.

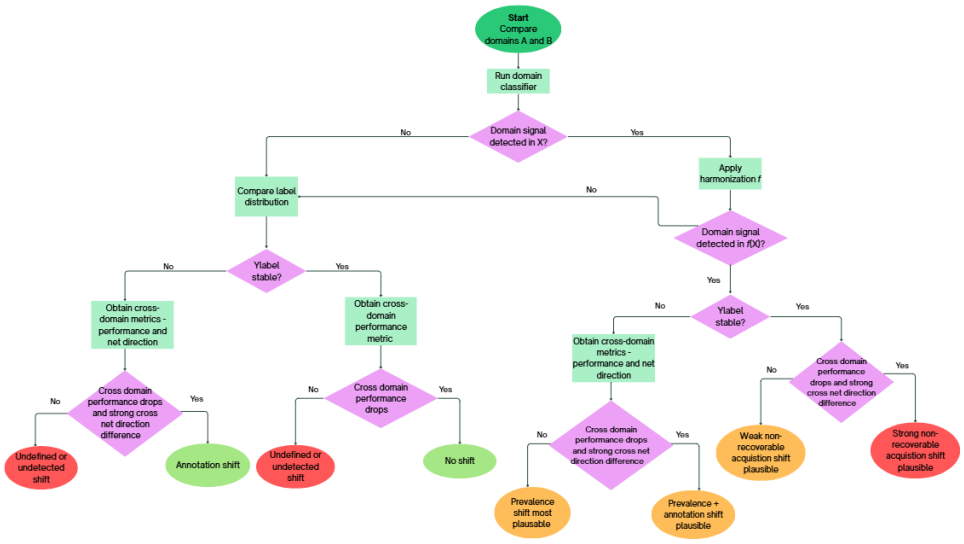


Figure 3: Diagnostic flowchart for interpreting cross-domain differences

6 Synthetic Experiments

To evaluate the diagnostic framework in a setting where the true source of shift is known, we construct a simple synthetic binary classification task. The goal of these experiments is not to reproduce the complexity of medical imaging data, but to test whether the proposed diagnostics behave as expected under controlled forms of dataset shift.

In this setup, the underlying target state Y_{true} is modelled as a continuous latent severity variable, while the observed label Y_{label} is binary. Annotation shift is instantiated as a change in the threshold used to convert Y_{true} into Y_{label}. The same mechanism extends naturally to ordinal settings, where systematic annotation differences may correspond to changes in one or more grading thresholds. This threshold-based construction represents a common and interpretable form of annotation shift, but it does not cover all possible changes in $P(Y_{\text{label}} \mid Y_{\text{true}})$. Other forms, such as changes in annotator noise, reliability, or class-specific confusion patterns, may lead to different empirical signatures.

Each synthetic sample consists of the latent continuous target state Y_{true}, a disease-relevant observed feature X_{rel}, a disease-irrelevant nuisance feature X_{irr}, and an observed binary label Y_{label}. The reference domain A is generated from

$$Y_{\text{true}} \sim \mathcal{N}(0, 1) \tag{4}$$

with the disease-relevant feature defined as

$$X_{\text{rel}} = Y_{\text{true}} + \epsilon_{\text{rel}}, \qquad \epsilon_{\text{rel}} \sim \mathcal{N}(0, 0.5^2), \tag{5}$$

and the disease-irrelevant feature generated independently as

$$X_{\text{irr}} \sim \mathcal{N}(0, 1). \tag{6}$$

The observed label is obtained by thresholding Y_{true} into two classes:

$$Y_{\text{label}} = \begin{cases} 0, & Y_{\text{true}} < 0, \\ 1, & Y_{\text{true}} \geq 0. \end{cases} \tag{7}$$

This setup represents a simplified diagnostic problem in which an underlying continuous target state is discretised into a binary label. The disease-relevant feature X_{rel} is an imperfect observation of the underlying target state, while X_{irr} represents disease-irrelevant nuisance variation.

Domain B is generated by modifying one or more components of the reference process. We first simulate cases where the shift is induced through the underlying target-state distribution or the labelling rule, rather than by directly modifying the observed features. These cases are intended to test whether the proposed diagnostics can distinguish annotation-related and prevalence-related mechanisms:

- **No shift:** Domain B is generated using the same parameters as domain A.

- **Annotation shift:** Only the labelling threshold is changed to $Y_{\text{label}} = \mathbb{I}(Y_{\text{true}} \geq -0.7)$, making domain B contain more positive cases.
- **Prevalence shift:** The mean of Y_{true} is increased to $Y_{\text{true}} \sim \mathcal{N}(0.7, 1)$, also making domain B contain more positive cases.
- **Combined prevalence and annotation shift:** Both the mean of Y_{true} and the labelling threshold are changed to $Y_{\text{true}} \sim \mathcal{N}(0.7, 1)$ and $Y_{\text{label}} = \mathbb{I}(Y_{\text{true}} \geq -0.7)$.

We then simulate acquisition-related variation by modifying the observed input representation X. These cases test whether input shifts can produce patterns that resemble annotation shift, and whether the effect depends on the type of information affected. We distinguish between task-irrelevant acquisition shift, where the changed nuisance feature should not affect the target task; task-irrelevant acquisition shift with label leakage, where a nuisance feature becomes predictive of Y_{label}; and disease-relevant acquisition shift, where the representation of Y_{true} in X is altered. This separates input shifts that are detectable but potentially harmless from shifts that can mislead the model or change the target-relevant representation.

- **Recoverable acquisition shift:** Only the nuisance feature is changed by shifting its mean to 2.0, so that $X_{\text{irr}} \sim \mathcal{N}(2.0, 1)$. It remains independent of both Y_{true} and Y_{label}.
- **Recoverable acquisition shift with label leakage:** The nuisance feature is constructed to contain information about the observed label, even though it does not represent the underlying target state:
  $X_{\text{irr}} = Y_{\text{label}} + \epsilon_{\text{irr}}$.
- **Non-recoverable disease-relevant acquisition shift:** The disease-relevant feature is changed. An offset is added to the disease-relevant feature, $X_{\text{rel}} = Y_{\text{true}} - 1 + \epsilon_{\text{rel}}$. This changes how the same underlying disease state is expressed in the observed input. Unlike nuisance variation, the shifted component is directly related to Y_{true}, so removing the domain signal would also alter task-relevant information. As a result, the acquisition shift cannot be corrected without affecting the target-relevant representation itself.

For each scenario, 5000 samples are generated for each domain. Both domains are split into training and test sets, stratified by the observed label. The diagnostics are estimated using simple linear models. The input-distribution diagnostic is estimated with a logistic-regression domain classifier trained to distinguish samples from domain A and domain B using the observed features $X = (X_{\text{rel}}, X_{\text{irr}})$; its held-out accuracy is used as the domain-classifier accuracy. The label-distribution diagnostic is estimated directly from the empirical

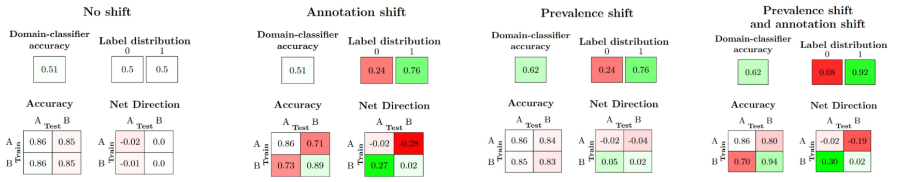
proportions of Y_{label} in each domain. For the bidirectional model-behaviour diagnostic, separate logistic-regression classifiers are trained to predict Y_{label} in domain A and domain B, respectively, and are evaluated in four directions:

$$A \to A, \quad A \to B, \quad B \to B, \quad B \to A.$$

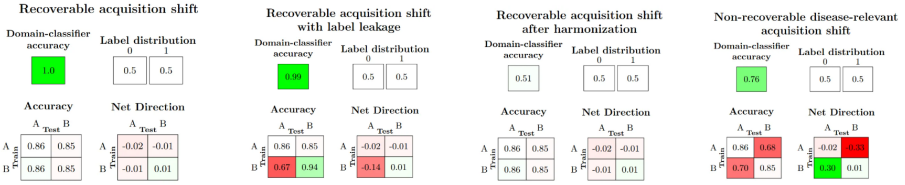
This allows within-domain performance to be compared with cross-domain transfer performance. Cross-domain performance is measured using accuracy, and directional behaviour is summarised using the net-direction metric defined above.

6.1 Synthetic Experiment Results

We next evaluate the diagnostic framework across the synthetic shift scenarios. Because the source of shift is known by construction, these experiments provide a controlled setting for comparing the expected diagnostic signatures of annotation, prevalence, and acquisition-related shifts.



(a) Target-state and labelling shifts.



(b) Acquisition-related shifts.

Figure 4: Diagnostic signatures of different shift mechanisms in the synthetic experiments. Domain-classifier accuracy, observed label proportions $P(Y_{\text{label}})$, cross-domain accuracy, and net direction are shown for each simulated shift case. Green and red shading indicate relative increases and decreases compared with the no-shift condition. Annotation shift is characterised by a reversal in net direction across transfer directions despite weak domain separability, whereas prevalence shift primarily changes observed label proportions and acquisition-related shifts are distinguished by stronger domain-classifier accuracy.

Figure 4(a) summarises the cases in which the underlying target-state or labelling process changes while the observed representation remains stable. The no-shift condition shows the expected reference pattern: near-chance domain

discrimination, similar label proportions, stable cross-domain accuracy, and near-zero net direction.

Annotation shift produces the clearest directional signature. Although the domain classifier remains near chance, indicating comparable observed inputs, the cross-domain models exhibit a reversal in net direction. The $A \to B$ model systematically over-predicts relative to the target-domain labels, whereas the $B \to A$ model under-predicts. This reflects the changed labelling threshold rather than an observable input-distribution shift.

Prevalence shift instead primarily changes the class proportions and introduces a detectable domain signal. Cross-domain predictive performance remains relatively preserved, while the directional effects are much smaller and more symmetric than under annotation shift. The combined prevalence-and-annotation scenario exhibits both patterns simultaneously: shifted class proportions together with an annotation-like reversal in net direction. The reversal is weaker than under pure annotation shift, suggesting that prevalence effects can partially offset the annotation-shift signal. Nevertheless, the pattern remains closer to annotation shift than to pure prevalence shift, where directional effects are weak and mostly symmetric.

Figure 4(b) summarises the acquisition-related shift scenarios, where the observed representation X is modified directly. The task-irrelevant acquisition shift without label leakage produces a strong domain signal but has little effect on cross-domain accuracy or net direction, indicating that detectable input differences do not necessarily affect task-relevant behaviour.

In contrast, the acquisition shift with label leakage substantially degrades cross-domain performance. Although the modified feature is intended to be disease-irrelevant, it becomes predictive of the observed label and therefore acts as a shortcut signal. This demonstrates how acquisition-related artefacts or annotation-related traces can mislead the model even when the underlying target-state relationship is unchanged.

After harmonisation, implemented by removing X_{irr} entirely from the feature representation, the domain signal decreases to near chance and the cross-domain metrics return close to their no-shift values. In this synthetic setting, harmonisation is possible because the nuisance feature responsible for the domain signal is known by construction and can be removed without affecting the disease-relevant representation X_{rel}.

Finally, disease-relevant acquisition shift produces both a persistent domain signal and a substantial drop in cross-domain accuracy. The resulting directional effects partially resemble annotation shift despite the labelling process remaining unchanged. This illustrates why input-distribution diagnostics are necessary when interpreting directional error patterns, since changes in the representation of (Y_{true}) within X can produce annotation-like behaviour even in the absence of annotation shift.

6.2 Robustness of Directional-error Patterns

The previous synthetic experiments suggested that annotation shift and prevalence shift produce qualitatively different behavioural patterns, particularly in the net-direction metric. However, these results were obtained under a single choice of annotation thresholds and target-state distributions. To assess whether the observed patterns reflect robust properties of the shift mechanisms rather than specific parameter choices, we next analyse how the net-direction metric changes under controlled variation of the domain-B annotation threshold, the domain-B target-state distribution, and their combination.

We focus on net direction because it was the most discriminative behavioural metric in the preceding experiments. Unlike overall performance measures, net direction captures whether cross-domain models fail systematically in opposite directions, making it particularly sensitive to annotation-related effects. Figure 5 compares the behaviour of this metric under pure annotation shift and pure prevalence shift.

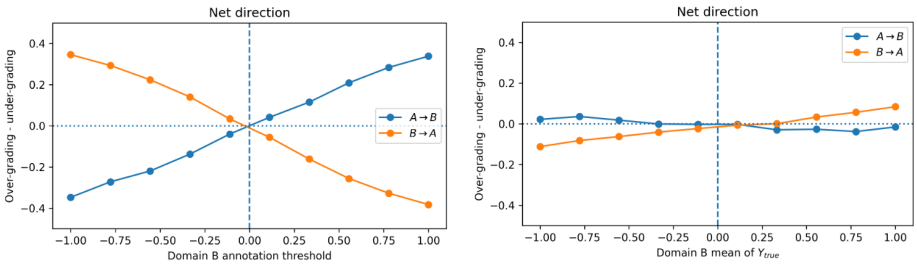


Figure 5: Left: Net direction under pure annotation shift. The annotation threshold in domain B is varied while the target-state distribution remains fixed. Right: Net direction under pure prevalence shift. The mean of Y_{true} in domain B is varied while the annotation threshold remains fixed. Annotation shift produces a strong reversal in directional errors across transfer directions, whereas prevalence shift produces a much weaker and more symmetric pattern.

In Figure 5, the two panels vary different parameters, but both are expressed on the same latent target-state scale. Since the reference domain uses $Y_{\text{true}} \sim \mathcal{N}(0, 1)$, a change of 0.5 in either the annotation threshold or the mean of Y_{true} corresponds to half a standard deviation on this scale. The effects should therefore be compared as changes of similar magnitude in the latent severity space, while keeping in mind that they act on different mechanisms: the annotation threshold changes $P(Y_{\text{label}} \mid Y_{\text{true}})$, whereas the mean shift changes the distribution of Y_{true}.

Under pure annotation shift, the directional-error pattern is strong and symmetric. When the domain-B threshold differs from the reference threshold, the two transfer directions move in opposite directions: the $A \to B$ model systematically over- or under-predicts relative to the target-domain labels, while the $B \to A$ model shows the reverse behaviour. The magnitude of the effect is also

large, with net direction reaching approximately -0.35 to 0.50 depending on the threshold difference.

In contrast, pure prevalence shift produces much smaller directional effects. Although changing the mean of Y_{true} alters the case mix in domain B, the net-direction values remain close to zero and do not exhibit the strong reversal pattern characteristic of annotation shift. This suggests that prevalence shift alone has a limited effect on directional behaviour relative to changes in the annotation threshold.

Figure 6 examines the combined prevalence-and-annotation-shift setting by varying the domain-B annotation threshold under different levels of prevalence shift in domain B. The resulting curves largely preserve the annotation-shift pattern: increasing the annotation threshold still produces opposite directional changes across transfer directions. This suggests that annotation shift remains distinguishable even when prevalence shift is also present.

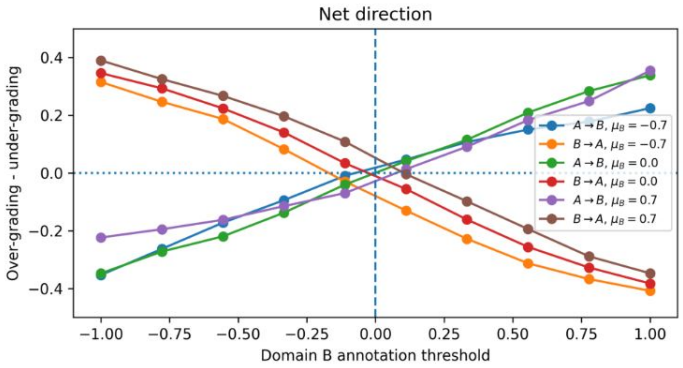


Figure 6: Net direction under combined prevalence and annotation shift. The annotation threshold in domain B is varied under different levels of prevalence shift in domain B. The annotation-shift pattern remains visible when threshold differences are sufficiently large, but small annotation shifts may overlap with prevalence effects.

At the same time, prevalence shift changes the baseline level of net direction by shifting the curves upward or downward. When the domain-B annotation threshold is sufficiently far from the reference threshold, the annotation-related reversal remains the dominant pattern. When the threshold remains close to the reference value, prevalence effects can partially mask or mimic the directional pattern. In this regime, the observed behaviour may be difficult to distinguish from prevalence-driven directional bias alone.

Together, these results support the use of net direction as a central behavioural diagnostic. In these simulations, pure annotation shift produces a stronger directional effect than pure prevalence shift for changes of comparable magnitude on the latent target-state scale, together with a systematic reversal in directional errors across transfer directions. When both mechanisms are

present, the annotation-shift signature can still remain visible, but its interpretation becomes more ambiguous when prevalence effects are sufficiently strong relative to the threshold difference.

7 Experiments on Osteoarthritis Radiographs

To evaluate the diagnostic framework in a more realistic medical-imaging setting, we use hip X-ray radiographs from the Osteoarthritis Initiative (OAI). The images are annotated with Kellgren–Lawrence (KL) osteoarthritis severity grades. Since the goal of these experiments is to study controlled forms of shift, the OAI data are used to construct paired domains in which specific shift mechanisms can be induced while keeping the underlying image set fixed.

The dataset is split into training, validation, and test sets containing 4708, 1008, and 1008 hip images, respectively. Both left and right hips are included. The split is performed at the subject level, so that images from the same subject do not appear in more than one of the training, validation, or test sets. This prevents subject-level data leakage between splits.

The original KL grades are converted into a binary classification task. In the reference domain, denoted domain A, the input radiographs are left unchanged and the original labels are binarized using the rule

$$Y_{\text{label}} = \mathbb{I}(\text{KL} \geq 2). \tag{8}$$

Thus, KL grades 0 and 1 are treated as negative cases, while KL grades 2 and above are treated as positive cases. Domain A therefore represents the clean reference domain. The resulting reference-domain label distribution is strongly imbalanced, with approximately 92% negative cases and 8% positive cases.

Domain B is constructed from the same underlying OAI images, but modified according to the shift mechanism being studied. This design allows differences in model behaviour to be attributed to the induced shift rather than to differences in the underlying patient population or image identities. Four shifted conditions are considered.

First, for the annotation-shift condition, the input radiographs are unchanged, but the binarization rule is changed to

$$Y_{\text{label}} = \mathbb{I}(\text{KL} \geq 1). \tag{9}$$

This makes domain B use a more inclusive labelling threshold than domain A. Since the images are unchanged, this condition is intended to isolate the effect of a changed annotation rule.

Second, for the prevalence-shift condition, the original binarization rule $\text{KL} \geq 2$ is kept, but 80% of the $\text{KL} = 0$ cases are removed from domain B. This changes the observed class composition without changing the labelling threshold for positive cases.

Third, for the acquisition-shift condition, the labels are kept unchanged but the images in domain B are degraded to simulate differences in image acquisition

quality, such as those that might arise from an older scanner. This is implemented by applying Gaussian blur with $\sigma = 1.4$ and reducing contrast with a contrast factor of 0.7. This condition is intended to introduce an input-domain shift while preserving the observed labels. We do not attempt to learn or apply a harmonising transformation $f(X)$ in this experiment. Therefore, this condition tests whether an uncorrected acquisition-related perturbation produces the diagnostic signatures expected for input shift, rather than whether the perturbation can be fully corrected.

Fourth, for the combined prevalence-and-annotation-shift condition, domain B uses both the modified annotation rule KL ≥ 1 and the prevalence-shift procedure described above. This condition tests whether the annotation-shift signature remains visible when case-mix differences are also present.

For each condition, convolutional neural-network classifiers are trained separately on domain A and domain B. The model architecture is a ResNet-18 pretrained on ImageNet and adapted for binary classification [22]. Images are resized to 160×160, and the final classification layer is replaced with a two-class output layer. Models are trained using cross-entropy loss and the AdamW optimizer with learning rate 10^{-4} and weight decay 10^{-4}. Training is performed for a maximum of 30 epochs with early stopping based on validation ROC-AUC, using a patience of 7 epochs and a minimum improvement threshold of 10^{-4}. The best checkpoint according to validation ROC-AUC is used for final evaluation.

The models are evaluated in the same bidirectional design used in the synthetic experiments. The within-domain evaluations $A \to A$ and $B \to B$ provide reference performance under each domain's own images and labelling rule. The cross-domain evaluations $A \to B$ and $B \to A$ test how model behaviour changes when the training and evaluation domains differ.

Three diagnostic quantities are reported. First, domain separability is measured using the held-out accuracy of a domain classifier trained to distinguish samples from domain A and domain B. This assesses whether the induced shift creates a detectable input-domain signal. Second, task performance is measured using AUROC in each evaluation direction. AUROC is used because it captures ranking performance and is less dependent on a fixed classification threshold [23]. Third, directional error behaviour is measured using net direction, which indicates whether a model tends to predict labels that are higher or lower than the observed target-domain labels.

This experimental design is useful because domains A and B are constructed from the same underlying image source, and in most conditions from the same image identities. As a result, observed differences in model behaviour are less likely to be explained by uncontrolled differences in patient populations or data collection sites. Instead, the comparison focuses on whether the induced changes in labelling rule, class composition, or image appearance produce the diagnostic signatures predicted by the framework.

7.1 Diagnostic Results on Osteoarthritis Radiographs

The OAI experiments show that the synthetic diagnostic patterns partly carry over to a more realistic imaging setting, but with weaker and less cleanly separated signals. Annotation shift still produces the clearest directional asymmetry, while prevalence shift mainly affects the observed label distribution and disease-relevant acquisition shift is most visible through domain-classifier accuracy. However, the strong class imbalance in the OA task makes the net-direction values harder to interpret than in the synthetic experiments.

Figure 7 reports domain-classifier accuracy, observed label proportions $P(Y_{\text{label}})$, cross-domain AUROC, and net direction for each induced shift condition. In the original domain A, the observed class proportions are 0.93/0.07, reflecting the strong class imbalance in the binarised OA task.

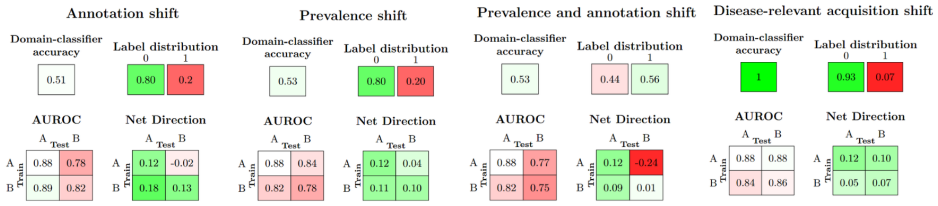


Figure 7: Diagnostic signatures of different shift mechanisms in the OAI experiments. Domain-classifier accuracy, observed label proportions $P(Y_{\text{label}})$, cross-domain AUROC, and net direction are shown for each induced shift condition. The annotation-shift condition produces the clearest directional asymmetry despite weak domain separability, consistent with the synthetic experiments. In contrast, prevalence shift produces smaller directional effects, and disease-relevant acquisition shift is characterised by strong domain-classifier accuracy without the annotation-like reversal in net direction. Compared with the synthetic setting, the OA experiments exhibit noisier and less symmetric patterns, reflecting the increased complexity of realistic medical-imaging data.

It is important to note that the baseline $A \to A$ net direction is already positive 0.12, indicating that even the reference model slightly over-predicts the positive class. This likely reflects class imbalance or thresholding effects in the OA data. Therefore, the cross-domain net direction values should be interpreted relative to this baseline, rather than as deviations from zero alone.

The annotation-shift scenario produces the clearest directional asymmetry. The domain-classifier performance remains near chance 0.51, indicating that the observed inputs remain comparable across domains. Relative to the positive $A \to A$ baseline net direction 0.12, the $A \to B$ model shifts downward to -0.02, whereas the $B \to A$ model shifts upward to 0.18. This pattern is consistent with the synthetic experiments: the changed labelling convention shifts the decision boundary between negative and positive cases, producing directional errors.

Prevalence shift produces a different signature. The domain classifier remains close to chance (0.53), and the observed class proportions are similar to

those in the annotation-shift condition. Cross-domain AUROC remains close to the within-domain baseline, and the net-direction values are small and similar across transfer directions. As in the synthetic experiments, prevalence shift does not produce the directional asymmetry characteristic of annotation shift.

The combined prevalence-and-annotation scenario exhibits both effects simultaneously. The class proportions shift to (0.44/0.56), while an annotation-like directional asymmetry is still visible: the $A \to B$ net direction decreases to -0.24, whereas the $B \to A$ net direction remains positive at 0.09. Compared with the pure annotation-shift condition, the asymmetry is less balanced across transfer directions, suggesting that prevalence effects alter the baseline directional behaviour. Nevertheless, the sign contrast between the two cross-domain directions is more consistent with annotation shift than with pure prevalence shift, where the net-direction values remain small and similar across directions.

Disease-relevant acquisition shift produces the strongest domain-classifier signal (1.0), indicating a large observable input difference, while the class proportions remain unchanged (0.93/0.07). In contrast to the annotation-shift scenarios, the cross-domain net-direction values remain positive in both transfer directions and do not show the sign reversal associated with annotation shift. Cross-domain AUROC remains similar to the within-domain values, suggesting that the perturbation is highly detectable as an input-domain shift but does not strongly disrupt ranking performance. This provides an important contrast: a substantial input-distribution difference alone is not sufficient to reproduce the annotation-like directional-error signature.

Overall, these results suggest that the proposed diagnostics remain informative in the OAI setting, but require more cautious interpretation than in the synthetic experiments. The annotation-shift and combined-shift conditions produce directional patterns that resemble the synthetic annotation-shift signature, whereas pure prevalence shift does not. At the same time, the baseline directional bias and strong class imbalance show that net direction should not be interpreted in isolation. Instead, the most useful evidence comes from the joint pattern across diagnostics: weak domain-classifier accuracy together with directional asymmetry supports an annotation-shift interpretation, while strong domain-classifier accuracy points toward acquisition-related differences.

8 Discussion

This work highlights the importance of distinguishing the underlying target state Y_{true} from the observed assigned label Y_{label} when reasoning about annotation shift. Changes in the observable relationship $P(Y_{\text{label}} \mid X)$ may reflect changes in labelling behaviour, but they may also arise from differences in prevalence, acquisition, or the way the underlying state is represented in the input. Defining annotation shift as a change in $P(Y_{\text{label}} \mid Y_{\text{true}})$ makes this distinction explicit and clarifies what kind of mechanism the framework is intended to diagnose.

The proposed framework should therefore be understood as a way to compare plausible explanations for observed cross-domain differences, rather than as a

definitive test for annotation shift. The main empirical pattern identified in the experiments is that systematic directional annotation differences can produce a reversal in cross-domain error direction: a model trained in one domain tends to over- or under-predict in the other domain, while the reverse transfer direction shows the opposite behaviour. This is useful because it provides information that is not captured by overall cross-domain performance alone.

At the same time, this diagnostic signature is most directly applicable to ordered and directional forms of annotation shift. The clearest example is a threshold shift, where one domain applies a stricter or more lenient criterion along an ordered severity scale. This is also the type of annotation shift induced in the experiments, including the relatively strong osteoarthritis shift from KL ≥ 2 to KL ≥ 1. Other forms of annotation shift, such as changes in annotator reliability, symmetric label noise, class-specific confusion, or non-monotone relabelling, may affect performance or calibration without producing a clear upward or downward error pattern. The directional-error diagnostic should therefore be interpreted as evidence for systematic directional annotation differences, rather than as a general detector of all possible annotation shifts.

The framework also depends on practical modelling choices. The input-distribution diagnostic depends on the domain classifier and on the representation of X used to compare domains. A near-chance domain classifier does not prove that the input distributions are identical; it only indicates that no domain signal was detected by the chosen classifier and representation. Similarly, the model-behaviour diagnostics depend on the predictive models used for the task. If the task model does not learn target-relevant structure well, then cross-domain error patterns may be difficult to interpret. In practice, the diagnostics should therefore be interpreted together with model performance, validation results, and domain knowledge about the data-generating process.

A related practical challenge is determining whether input differences can be corrected without removing information about the underlying target state. This would require a transformation $f(X)$ that reduces domain-specific acquisition variation while preserving target-relevant information. In practice, this can only be checked indirectly, for example by verifying that the domain-classifier signal decreases after applying $f(X)$, while within-domain task performance or other measures of target-relevant information remain comparable. If acquisition differences affect disease-relevant image regions, such a transformation may be difficult to define or validate.

Studying annotation shift empirically is difficult because the true source of shift is usually unknown and several mechanisms may occur at the same time. For this reason, the synthetic experiments and constructed OAI domains were used to create controlled settings in which specific shift mechanisms could be isolated. This makes the diagnostic signatures easier to interpret, but it also simplifies the deployment setting. In a real multi-institution osteoarthritis setting, institutions may differ simultaneously in imaging protocols, patient populations, disease severity distributions, and grading conventions. The framework could still be useful as an exploratory diagnostic in such settings, but the interpreta-

tion would be less direct than in the controlled experiments.

Prevalence shift is particularly important in this respect. A prevalence shift involving visually obvious disease states may produce a detectable input-domain signal, but this is not guaranteed. Case-mix differences may affect mainly borderline cases, may be small relative to other sources of image variability, or may be difficult for the chosen domain classifier to detect. Prevalence and annotation effects may also partly mask or offset each other in the observed label distribution. The robustness experiments illustrate this issue: when annotation and prevalence shifts are combined, the annotation-shift pattern can remain visible, but it becomes less cleanly separated when prevalence effects are strong relative to the threshold difference.

More broadly, annotation shift is difficult to identify from observed data alone because Y_{true} is usually unobserved. The proposed diagnostics can show that observed patterns are consistent with annotation shift, but they cannot prove that $P(Y_{\text{label}} \mid Y_{\text{true}})$ has changed without additional information. Stronger evidence would require either a proxy for the underlying target state, such as expert consensus labels, independent clinical measurements, or longitudinal outcomes, or assumptions that link the observed inputs and labels to Y_{true}. For example, longitudinal outcomes may provide information about disease progression or clinical state that is less dependent on the original grading convention.

Future work should therefore move in two directions. First, the framework should be evaluated on independent multi-institution datasets where acquisition conditions, patient populations, and annotation practices may all differ. Second, additional experiments should study mixed, weaker, and non-threshold forms of annotation shift, including changes in annotator reliability, class-dependent noise, and non-monotone labelling differences. This would help determine which diagnostic signatures are specific to threshold-like annotation shifts and which generalise to broader changes in $P(Y_{\text{label}} \mid Y_{\text{true}})$.

9 Conclusion

This thesis introduced a diagnostic framework for assessing whether observed cross-domain differences are consistent with annotation shift. By distinguishing the underlying target state Y_{true} from the observed assigned label Y_{label}, and by defining annotation shift as a change in $P(Y_{\text{label}} \mid Y_{\text{true}})$, the framework separates changes in labelling behaviour from prevalence- and acquisition-related effects.

Across both synthetic and osteoarthritis experiments, different shift mechanisms produced different diagnostic patterns. Systematic threshold-based annotation shift was associated with directional asymmetries in bidirectional model evaluation, whereas prevalence and acquisition shifts produced different combinations of label-distribution, input-distribution, and model-behaviour signatures.

These results suggest that annotation shift should not be inferred from changes in $P(Y_{\text{label}} \mid X)$ alone. Instead, it should be assessed by comparing

multiple observable diagnostics and by considering alternative explanations for cross-domain differences. The proposed framework provides a step toward such interpretation, while highlighting that stronger evidence for annotation shift ultimately requires either additional information about the underlying target state or assumptions linking Y_{true}, X, and Y_{label}.

References

[1] Matthew S. Harkey, Kerry E. Costello, Bella Mehta, Chunyi Wen, Anne-Marie Malfait, Henning Madry, and Brooke E. Patterson. Artificial intelligence in osteoarthritis research: Summary of the 2025 oarsi pre-congress workshop. *Osteoarthritis and Cartilage Open*, 7(4):100687, 2025. doi: 10.1016/j.ocarto.2025.100687. URL https://www.sciencedirect.com/science/article/pii/S2665913125001232.

[2] Suruchi Kumari and Pravendra Singh. Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *Computers in Biology and Medicine*, 170:107912, 2024. doi: 10.1016/j.compbiomed.2023.107912. URL https://www.sciencedirect.com/science/article/pii/S001048252301377X.

[3] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4396–4415, 2023. doi: 10.1109/TPAMI.2022.3195549. URL https://doi.org/10.1109/TPAMI.2022.3195549.

[4] Mélanie Roschewitz, Raghav Mehta, Charles Jones, and Ben Glocker. Automatic dataset shift identification to support safe deployment of medical imaging ai. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2025*, volume 15966 of *Lecture Notes in Computer Science*, pages 67–76. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-04981-0$_7$.*URL*

[5] Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11:3673, 2020. doi: 10.1038/s41467-020-17478-w. URL https://www.nature.com/articles/s41467-020-17478-w.

[6] Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G. Dy. Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3):291–327, 2014. doi: 10.1007/s10994-013-5412-1. URL https://doi.org/10.1007/s10994-013-5412-1.

[7] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2009. ISBN 9780262170055.

[8] Jose G. Moreno-Torres, Troy Raeder, Rocio Alaiz-Rodriguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012. doi: 10.1016/j.patcog.2011.06.019.

[9] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C*, 28(1):20–28, 1979. doi: 10.2307/2346806.

[10] R. D. Altman and G. E. Gold. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthritis and Cartilage*, 15 (Suppl A):A1–A56, 2007. doi: 10.1016/j.joca.2006.11.009. URL `https://www.sciencedirect.com/science/article/pii/S1063458406003281`.

[11] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 2020. doi: 10.48550/arXiv.2004.05785.

[12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[13] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

[14] Zachary C. Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130, 2018.

[15] Felipe Maia Polo, Rafael Izbicki, Evanildo Gomes Lacerda Jr, Juan Pablo Ibieta-Jimenez, and Renato Vicente. A unified framework for dataset shift diagnostics. *Information Sciences*, 649:119612, 2023. doi: 10.1016/j.ins.2023.119612. URL `https://arxiv.org/abs/2205.08340`.

[16] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018. doi: 10.1371/journal.pmed.1002683.

[17] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 136–155, 2020. URL `https://proceedings.mlr.press/v121/cohen20a.html`.

[18] Eduardo H. P. Pooch, Pedro L. Ballester, and Rodrigo C. Barros. Can we trust deep learning models diagnosis? the impact of domain shift in chest

radiograph classification. *arXiv preprint arXiv:1909.01940*, 2019. URL `https://arxiv.org/abs/1909.01940`.

[19] Suruchi Kumari and Pravendra Singh. Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *Archives of Computational Methods in Engineering*, 31(2): 1307–1330, 2024. doi: 10.1007/s11831-023-09962-0.

[20] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446.

[21] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622, 2008. doi: 10.1145/1401890.1401965.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2016.

[23] James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1): 29–36, 1982. doi: 10.1148/radiology.143.1.7063747.