Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Scalable Drift Monitoring in Medical Imaging AI

Jameson Merkow
Microsoft Health and Life Sciences (HLS)
Felix J. Dorfner
Massachusetts General Hospital
Xiyu Yang
Massachusetts General Hospital
Alexander Ersoy
Microsoft Health and Life Sciences (HLS)
Giridhar Dasegowda
Massachusetts General Hospital
Mannudeep Kalra
Massachusetts General Hospital
Matthew P. Lungren
Microsoft Health and Life Sciences (HLS)
Christopher P. Bridge
cbridge@mgh.harvard.edu

Massachusetts General Hospital
Ivan Tarapov
Microsoft Health and Life Sciences (HLS)

Article

Keywords:

Posted Date: February 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8744472/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. This work was partially supported by Microsoft, a global technology company that offers both paid and free products, including AI monitoring solutions which are potentially associated with patents.

Version of Record: A version of this preprint was published at npj Digital Medicine on October 27th, 2026. See the published version at https://doi.org/10.1038/s41746-026-03301-1.

Scalable Drift Monitoring in Medical Imaging AI

Jameson Merkow[1,*,†], Felix J. Dorfner[2,3,†], Xiyu Yang[2,3], Alexander Ersoy[1], Giridhar Dasegowda[2,3], Mannudeep Kalra[2,3], Matthew P. Lungren[1], Christopher P. Bridge[2,3,‡,+], and Ivan Tarapov[1,‡]

[1]Microsoft Health and Life Sciences (HLS), Redmond, WA, USA
[2]Department of Radiology, Massachusetts General Hospital, Boston, MA, USA
[3]Harvard Medical School, Boston, MA, USA
[*]jameson.merkow@microsoft.com
[+]corresponding author: cbridge@mgh.harvard.edu
[†]these authors contributed equally to this work
[‡]co senior authors

ABSTRACT

The integration of artificial intelligence (AI) into clinical medical imaging has advanced diagnostics but also poses challenges in maintaining long-term reliability as input data change. To address this, we present MMC+, a modular framework for scalable drift monitoring that builds on our CheXstray framework for real-time drift detection using multi-modal data concordance (MMC). MMC+ provides a flexible architecture that enables cost-effective continuous performance monitoring in real-world healthcare settings. The framework's capabilities include aggregating information from multiple data streams from statistical tests or distance metrics, leveraging foundation model embeddings for generalization across sites without requiring site-specific training, and quantifying uncertainty to better communicate drift significance. Validated with data from Massachusetts General Hospital during the COVID-19 pandemic, MMC+ effectively detects significant data shifts that correlate with performance changes and indicates when AI systems may deviate from acceptable performance bounds. The code for MMC+ is publicly available (Link: https://github.com/FJDorfner/MedImaging-ModelDriftMonitoring/tree/MMC_plus), promoting broader adoption of AI solutions in clinical settings. Our framework significantly advances drift detection while promoting transparency and collaboration in clinical AI research.

1 Introduction

The role of artificial intelligence (AI) in medical imaging has expanded significantly in recent years as evidenced by the increasing number of academic publications and the rising adoption of AI applications in clinical settings[1–5]. To ensure the safe and effective use of these AI algorithms, researchers and clinicians are developing a broad set of best practices for evaluation and deployment. These practices include "day 0" validation, which typically involves evaluation using a curated retrospective dataset to assess AI model performance prior to deployment[6,7]. As the field matures and clinical AI governance practices are being developed, there is also growing recognition of the need for continuous, explainable monitoring tools that mitigate the risks of prediction drift, feature drift, and input data shifts after a model is deployed[8–10]. Drift arises commonly in imaging due to factors including changes in equipment, clinical protocols, and patient demographics, and recent cases studies have highlighted the risk that such drift can lead to significant degradation of model accuracy following deployment[11–13].

Notably, the FDA has underscored the importance of robust, real-time monitoring in AI-enabled medical devices, mandating adaptive oversight to ensure these systems remain safe and effective as clinical environments evolve[14]. In its recent guidance on AI-enabled device software functions, the FDA further emphasizes the necessity for comprehensive post-market surveillance strategies, including continuous performance tracking, early detection of data and feature drifts, and dynamic recalibration protocols.

Despite the recognized importance of continuous and explainable monitoring, many organizations rely on direct user reporting of model performance[15,16]. While this offers an effective means aims to assess whether the model's predictions remain accurate, it frequently fails to identify the root cause of performance changes such as changes in the input data. Moreover, direct performance monitoring requires timely access to ground truth labels, which are often provided manually, and as such monitoring continuously in this manner remains cost-prohibitive and may diminish the efficiency benefits of AI. If instead checks are performed periodically, the result may be delayed detection and response to problems. Furthermore, in some situations, including prognostic models, immediate access to ground truth is impossible and other methods of model monitoring are therefore a necessity.

A promising approach to overcoming the limitations of direct performance monitoring involves monitoring for the changes in the statistical properties of input data streams and model outputs using fully automated methods. Previous studies and methodologies have made significant strides in the monitoring of machine learning models in this way, however these primarily focus on other domains, for example the monitoring of Electronic Health Record (EHR) data[17,18] and lack support for imaging data. Supporting imaging data often requires generating image embeddings to create low-dimensional summaries of images for the purpose of drift detection, with traditional approaches like Variational Autoencoders (VAEs)[19] posing several challenges. VAEs require extensive site-specific training, lack the expressiveness to capture the complexity of medical images, and may necessitate separate models for different modalities or locations. This increases operational complexity and computational burden, limiting scalability and adaptability across institutions with diverse imaging protocols.

Our prior work with CheXstray[20] addresses this gap by developing a framework to support the automated monitoring of medical imaging AI models designed to include comprehensive image data analysis. CheXstray introduced a novel metric called multi-modal concordance (MMC), which aggregates metrics from various sources to assess overall data drift, and crucially does not assume access to the ground truth labels needed to monitor model performance. In an experimental setting, MMC demonstrated a strong correlation with model performance when artificial drift was induced within a datastream. However, recent research has shown that real-world data is more complex, and relying solely on surrogate metrics like MMC for performance monitoring can be ineffective[21].

Building on our previous conference paper[20], this work presents several key contributions to the field of AI model monitoring for medical imaging. We present MMC+, an extension that solidifies the framework and includes design principles for effective drift monitoring systems. The contributions of this manuscript include (1) identifying essential properties that drift monitoring components should possess for robust deployment across diverse clinical settings; (2) demonstrating how foundation models provide image representations that satisfy scalability requirements; (3) conducting comprehensive real-world validation during COVID-19 to evaluate these design principles; (4) analyzing the relationship between drift signals and model performance to understand when monitoring indicates the need for intervention; and (5) presenting an operational approach showing how characterized baseline ranges guide decisions about performance audits versus system stability. These insights provide guidance for architecting drift monitoring systems in medical imaging AI.

The implications of these findings are significant for the field of medical imaging AI, where data shifts can occur due to a variety of factors, such as changes in patient demographics, disease prevalence, and imaging protocols. By leveraging data drift monitoring, we can gain insight into different factors that drive change in model performance, helping us identify root causes, addressing gaps in many current monitoring solutions. This approach aligns with the need for a systematic, real-time performance monitoring framework that can adapt to the unique challenges posed by medical imaging data, ultimately supporting the broader adoption and integration of AI solutions in healthcare settings and simplify regulatory compliance by providing traceable, auditable performance metrics that align with FDA requirements.

2 Results

Our proposed MMC+ metric, shown in Figure 1, quantifies the dissimilarity between data in a given temporal window and a well-characterized pre-determined reference window, on which model accuracy is known (see Section 4). The higher the value of MMC+ in a given temporal window, the more dissimilar the data in that window is from the reference set. MMC+ aggregates three sources of information to give a unified dissimilarity score: image metadata, image appearance (using embeddings from the MedImageInsight foundation model[22]), and model outputs.

We conducted various experiments to demonstrate the use of MMC+ on a representative AI model operating on a large, real-world, consecutive dataset. The AI model is a DenseNet-121 convolutional neural network (CNN) trained to detect nine findings (atelectasis, cardiomegaly, edema, lung opacity, pleural effusion, pleural other, pneumonia, pneumothorax, and support devices) independently in an input radiograph following the Chexpert convention[23], see Section 4.3 for more details. The dataset consists of 90,581 chest radiograph studies from 60,062 patients, collected from Massachusetts General Hospital (MGH) between July 2019 and June 2021, representing all chest radiographs acquired during that time window at MGH. The data capture variability across clinical settings, including a mix of inpatient, emergency, and outpatient studies. We divided this dataset into training (July–September 2019), reference (October–December 2019), and test (January 2020–June 2021) sets to mirror a real-world model development and deployment process. Labels for the nine findings were generated from the accompanying radiology reports using a large language model. This dataset setup allowed us to evaluate MMC+ across different temporal and clinical shifts (see Section 4.2).

For all the experiments presented in the following sections, we used a window duration τ of 30 days and a window stride δ of 1 day (see Section 4.1). We chose this window duration as a balance between minimizing lag time and addressing the noise and other issues that arise with smaller sample sizes, even though this choice can introduce some delay in drift detection. The number of resampling iterations, N_s, was set to 20 (see Section 4.1).

2.1 MMC+ Results on Real-World Data

To evaluate the effectiveness of the MMC+ framework in detecting data drift, we applied MMC+ to a real-world dataset of chest radiographs collected over an 18 month period that includes the onset of the COVID-19 pandemic. This experiment aims to simulate a realistic scenario where an AI model faces significant shifts in data distribution, allowing us to assess how MMC+ responds to such changes and whether it can serve as a reliable indicator for initiating performance audits or model updates.

Figure 2 shows the MMC+ over time starting from November 1st, 2019, which marks the end of the first complete 30 day window within the reference set, until July 1st, 2021. The MMC+ metric remains steady for approximately the first 2 months but increases rapidly in mid March 2020. This also coincides with the declaration of the state of emergency in the state of Massachusetts on March 10, 2020 due to the COVID-19 pandemic[1]. Thereafter, MMC+ fluctuates but always remains high, suggesting that the data never returns to its initial characteristics following the onset of the pandemic. From Figure 2 we see how the following individual component groups contribute to MMC+: metadata, image appearance, and model output, see Section 4.4 for more details. All of these component groups show a similar trends, suggesting all groups detect the dramatic changes.

Next, we examine the model's performance in Figure 3, which depicts the AUROC for the individual findings across the same time period. We notice that nearly all performance indicators show some change starting in mid-March 2020, though they vary in scale and direction.

Starting with the most significant changes, we see that the performance for cardiomegaly drops significantly and remains low throughout the period. In contrast, "pleural other" initially shows an increase in performance but then dips around June 2020. We also observe that the performance for lung opacity detection improves starting in May 2020, despite significant changes in the data.

The remaining differences in performance are more subtle but still apparent. We see minor increases in performance for atelectasis and pneumonia. Findings such as edema, pleural effusion, support devices, and both averages (macro and micro) exhibit performance swings in both directions during this time period, indicating instability in model predictions.

Regarding pneumothorax performance, we observe large fluctuations within the performance curve. However, since the performance of predicting this finding also varied widely in the reference set, it is challenging to determine if these changes are connected to data drift or are due to inherent variability in the model's predictions for this condition.

These observations highlight the complex relationship between data drift and model performance. While MMC+ effectively detects shifts in the data distribution, the impact of this drift on model performance is not uniform across different findings.

2.2 Relationship Between MMC+ and Performance

To assess the relationship between the MMC+ metric and performance consistency, we evaluated how MMC+ values correspond to the classifier's normalized AUROC across the nine findings, as well as their micro and macro averages. Performance variability was observed across all datastreams, which prompted the calculation of "normal" fluctuations based on AUROC values from the reference set (see Section 4.1 for more details). These values were collected in 30 day windows to calculate the mean and standard deviation, which were subsequently used to normalize AUROC during the test period.

Each plot in Figure 4 presents a scatter plot where the x-axis represents the MMC+ value and the y-axis corresponds to the normalized AUROC for a different label. Data points are color-coded based on whether they occur before or after the onset of the pandemic (defined as the March 10 declaration of emergency in the state of Massachusetts), with the edges of each scatter plot displaying a kernel density estimate (KDE) of the two groups to highlight any separation.

A clear pattern emerges from these figures: when MMC+ values are within a certain range (less than 10 in this case), the normalized AUROC remains consistently within 3 standard deviations of the reference set mean. However, when MMC+ exceeds this range, the likelihood of the normalized AUROC remaining within 3 standard deviations of the reference mean diminishes significantly. Although performance does not universally decline—performance for some labels, such as atelectasis, even improves in certain instances—the variability in performance increases notably. This indicates that while MMC+ remains within its normal range, performance consistency is maintained, but outside this range, the model's reliability becomes less predictable. In other words, high MMC+ is a sensitive but not specific indicator of performance shifts.

These observations are further supported by summary statistics presented in Table 1. Across the board, we observe greater variation in performance when MMC+ is outside its normal range. Furthermore, in all cases except for Pneumothorax, which varies a great deal in the reference set, there is an increased proportion of AUROC values that deviate more than 3 standard deviations from the reference mean, underscoring the importance of maintaining MMC+ within its optimal range for consistent model performance.

Label	MMC+ < 10	MMC+ >= 10
Atelectasis	0.907	0.385
Cardiomegaly	0.711	0.049
Edema	0.845	0.810
Lung Opacity	0.825	0.100
Pleural Other	0.979	0.487
Pleural Effusion	1.000	0.885
Pneumonia	0.907	0.507
Pneumothorax	1.000	1.000
Support Devices	1.000	0.836

Table 1. Rates at which AUROC remains within 3 standard deviations of the reference set AUROC for various conditions across two MMC+ ranges (<10 and >=10), highlighting how the model's performance varies for different conditions as MMC+ increases.

2.3 Utility of Metadata

Figure 5 demonstrates the importance of incorporating metadata attributes sourced from both DICOM files and the Radiology Information System (RIS) into the monitoring framework (see Section 4.4). Specifically, the figure highlights a significant shift in the distribution of the "View Position" attribute for X-ray acquisitions starting around June 2020. Initially, less than 20% of the images were captured using the antero-posterior (AP) projection, but by mid-2020, this proportion had increased to over 40%. The heart and mediastinum appears larger on AP images than on PA images due to the increased distance to the image receptor and beam divergence, which negatively impacts the diagnostic quality of these images for cardiomegaly, an enlarged heart. Figure 5 shows that this shift in the proportion of the view position is accompanied by a notable decline in the model's performance for detecting cardiomegaly.

2.4 Comparison to p-value Based Methods

Several other methods for monitoring data drift track the p-value of a statistical hypothesis test between the data within a detection window and that in the reference set. For example, Kore et al.[21] monitor data drift using hypothesis test p-values. In Figure 5, we followed this approach to explore such behavior on our dataset.

Monitoring p-values without considering the associated effect size, especially when the number of samples can be large, is sensitive to distribution shifts that are likely inconsequential in practice, but statistically significant. In real-world data, this results in plots that appear volatile and difficult to interpret. By contrast, we find that distribution similarity methods accompanied by appropriate uncertainty bounds give values that better lend themselves to intuitive visualization of data drift. Unlike volatile p-values, methods that directly quantify distributional similarity offer more stable measures for drift quantification (see Section 4.1).

2.5 Monitoring at Separate Points of Care

Hospitals typically use a wide range of X-ray devices at different points of care, such as operating rooms, emergency departments, and outpatient clinics. Each of these points of care may experience data drift events differently, as they often deal with distinct patient populations, imaging protocols, and workflows. Monitoring drift separately at each point of care is therefore crucial to gaining a complete and nuanced understanding of drift events. Figure 6 illustrates the behavior of the MMC+ metric for two different points of care within the Massachusetts General Hospital system.

Figure 6 a) shows the MMC+ metric for the emergency room, where a pronounced drift event occurs around the onset of the COVID-19 pandemic. This drift mirrors the broader trend observed in the overall MMC+, with the onset of the pandemic causing a sharp change in the datastream. By contrast, Figure 6 b) shows the MMC+ metric for an operating room in the ambulatory care center. Here, the drift pattern is noticeably different. Although the data changes around the same time, the drift is marked by a broadening of the uncertainty bounds indicating a reduction in the number of studies performed. Additionally, a 50-day gap in data collection follows, after which the data resumes without a sustained drift. This also showcases the importance of uncertainty bounds.

2.6 Re-calibration After Drift

One of the operational challenges in deploying AI models in clinical settings is the inevitability of data drift. Over time, the input data characteristics will shift, leading to changes in model performance. Eventually, these shifts stabilize, creating a new "normal" state. In such scenarios, consider a situation where data at a specific site has drifted, prompting an audit. Following

[1] https://www.mass.gov/info-details/covid-19-state-of-emergency

this audit, necessary IT operations or model modifications are implemented to ensure that the model's performance returns to an acceptable range. However, to ensure continued monitoring effectiveness in this new data environment, it becomes essential to re-calibrate the MMC+ metric to track changes from this new reference set. Re-calibration of the MMC+ metric involves selecting a new reference set and recalculating the relevant weights and normalization factors (see Section 4.1).

This re-calibration is achieved by creating a new reference set using data from the audited period and recalculating the relevant weights and normalization factors. Our analysis as shown in Figure 6 c) demonstrates that this approach is effective, as it successfully adjusts the MMC+ metric back into its acceptable range, thereby restoring model stability and ensuring consistent performance in the newly stabilized data environment.

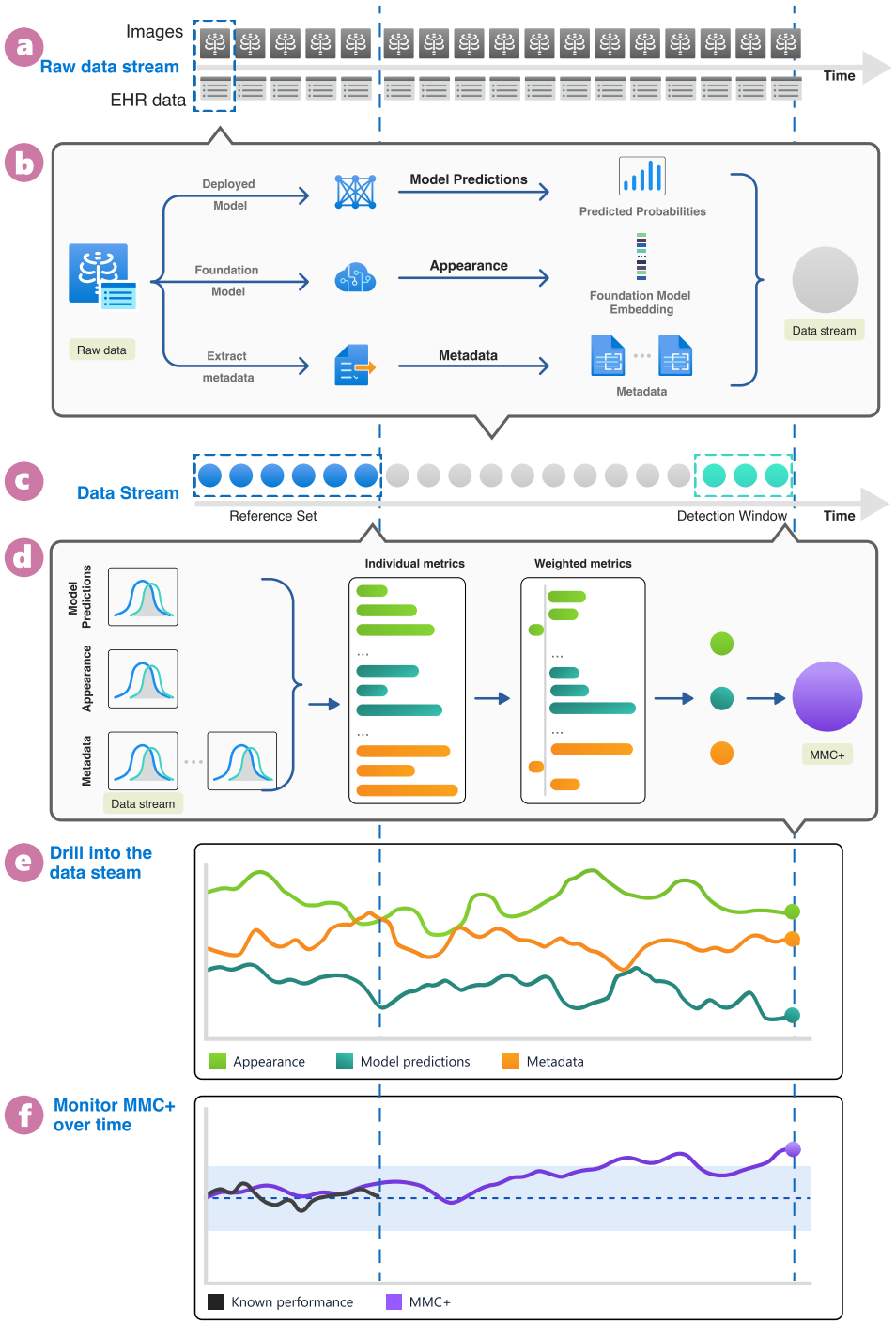


Figure 1. Illustration of the CheXStray framework for calculating the MMC+ metric. MMC+ is calculated by comparing a reference set with a detection window from ongoing data streams (panel **a**) to assess distribution similarity. High MMC+ values indicate potential deviations from normal performance thresholds. Separate processes process the input data stream (panel **b**) to create derived data (panel **c**) for each of the three components of MMC: Appearance, Model Predictions, and Metadata. Each individual component's distribution is evaluated against the reference set to calculate a similarity metric, which is then normalized and weighted (panels **d** and **e**) before being aggregated into the MMC+ metric. Panel **f** presents an example of MMC+ over time, displaying its relationship to known performance values and acceptable performance bands.

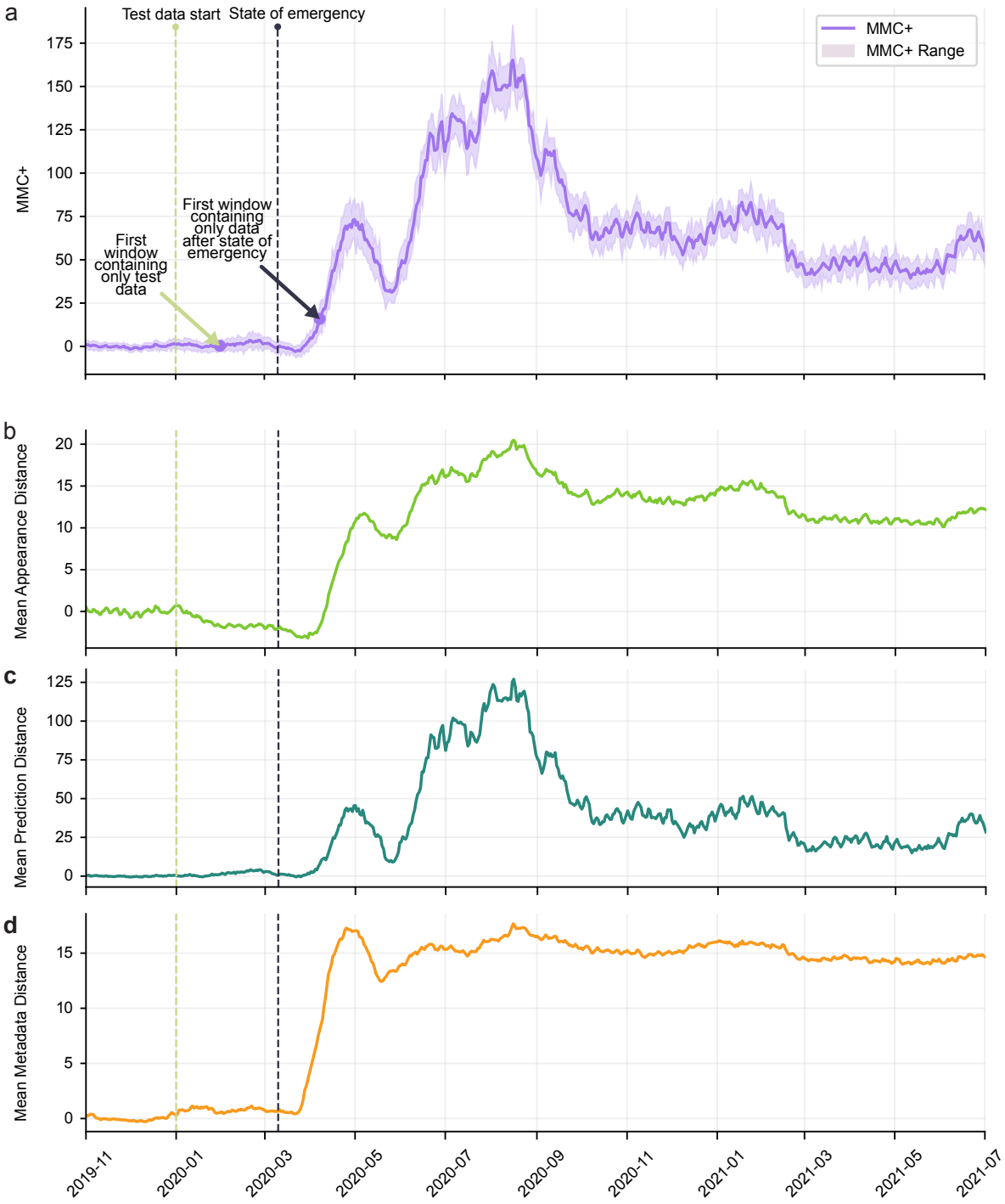


Figure 2. a: Plot of the evolution of MMC+ over time from 2019-11 to 2021-07. The purple line depicts the weighted MMC+ value and the purple shaded area indicates the range or variability around the MMC+, providing the uncertainty in the MMC+ measurement during this time period. Two dashed vertical lines are shown: the first marks the start of the test data, and the second represents March 10, 2021, the day Massachusetts declared a state of emergency. Two arrows highlight the first windows where all data is sourced from either the test data or post-state-of-emergency period, a consequence of the window duration. **b-d:** The MMC+ metric is composed of three high level components that capture data drift: the appearance metric (b) represents changes in the appearance of medical images as encoded by the MedImageInsight foundation model, the model prediction metric (c) captures shifts in model predictions from a trained classifier providing changes in the model's output over time, and the metadata metric (d) measures variations in the extracted from the DICOM files and RIS systems.

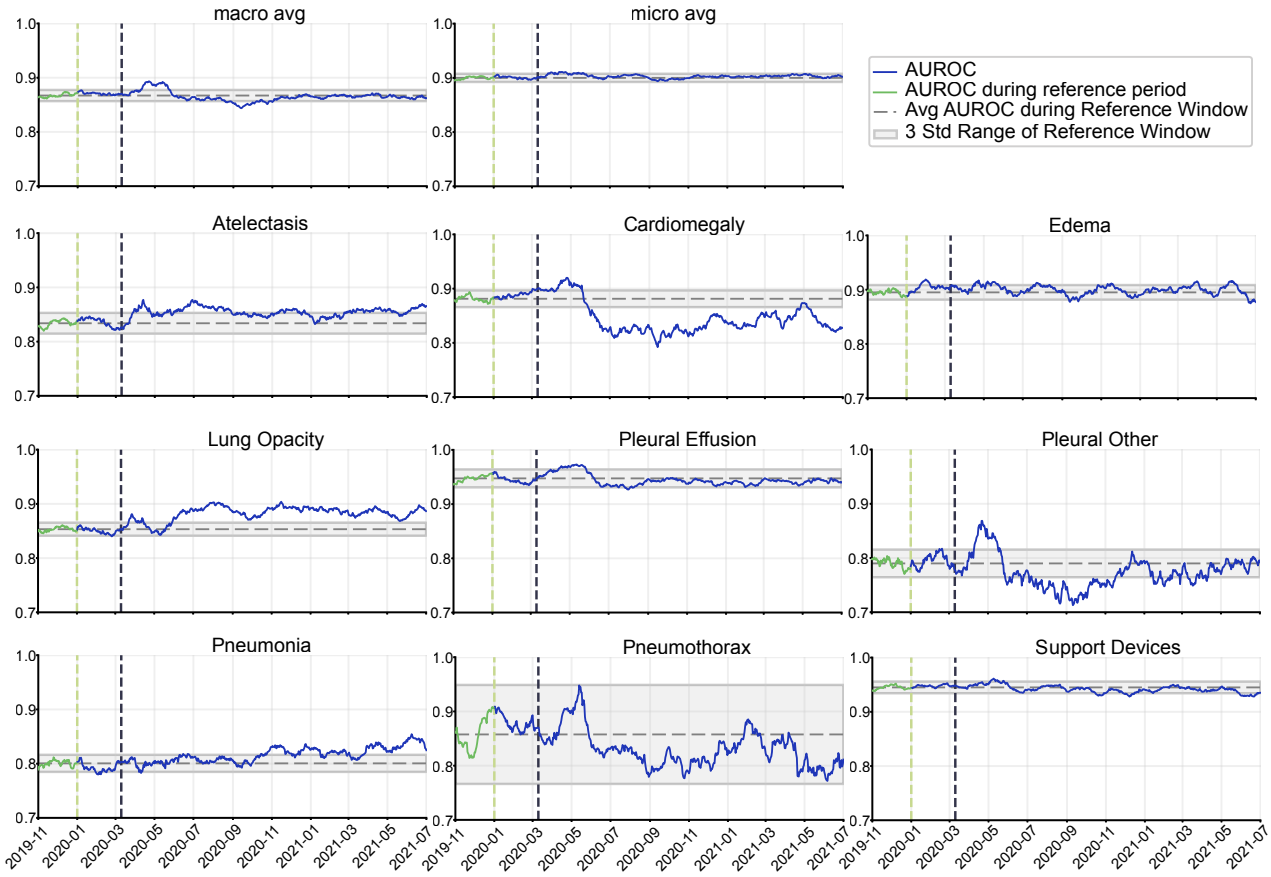


Figure 3. Plot of model performance (AUROC) for the nine findings and the macro/micro averages. AUROC is marked in blue, with the average performance during the reference period shown as a solid grey line and the reference period's 3 standard deviation range as a grey shaded area for each finding.

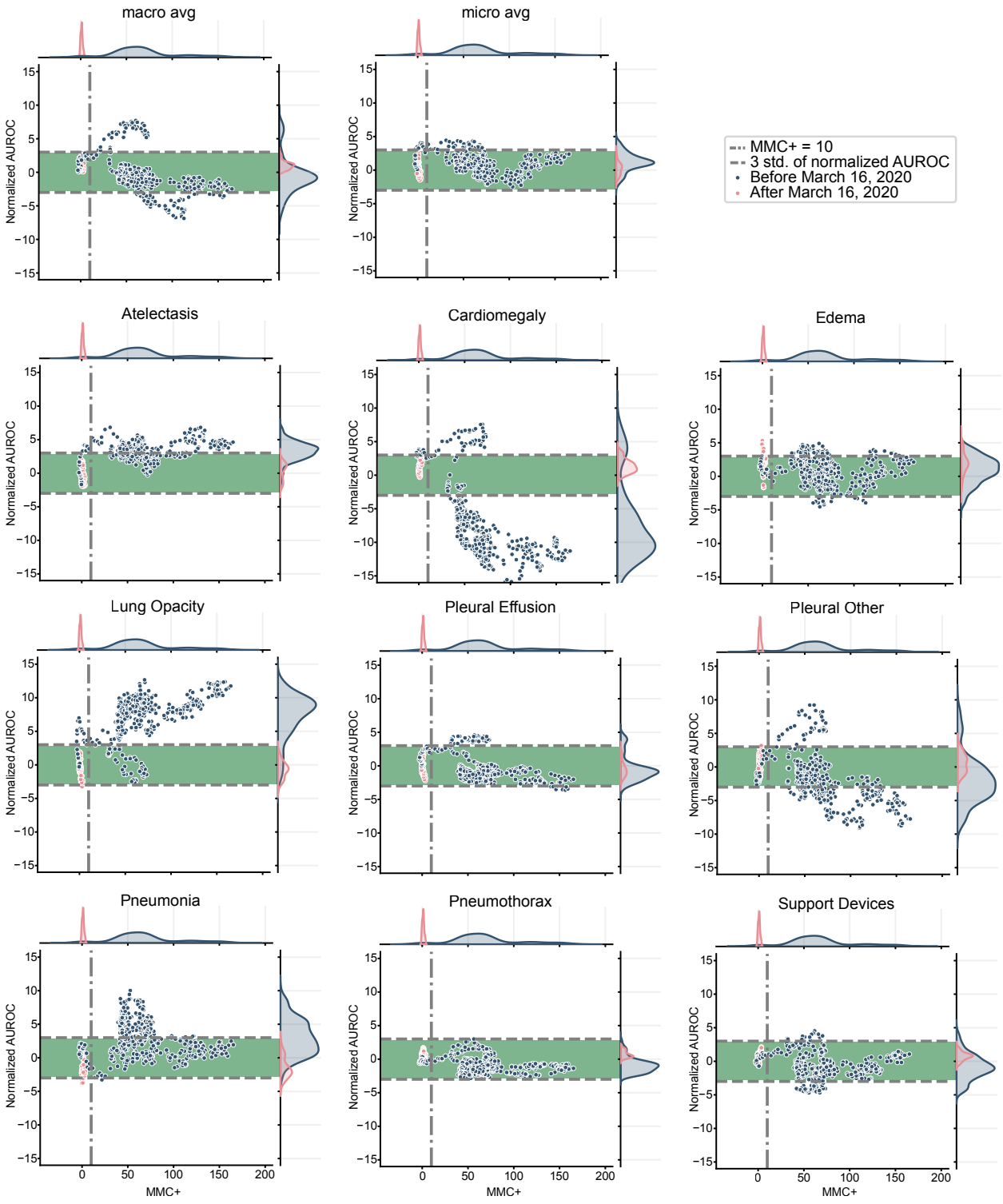


Figure 4. Relationship between MMC+ and normalized AUROC (standard deviation of AUROC) for each finding. In each plot, the pink points represent test set data before March 10, 2020, while the blue points represent data after this date. The horizontal dotted lines mark ± 3 standard deviations of the AUROC, and the vertical line at MMC+ $= 10$ represents the threshold where drift is observed. Above each scatter plot, KDE plots for the MMC+ and normalized AUROC are displayed, showing the distribution of test data pre- and post-March 10 (pink for pre, blue for post). These plots highlight the separation between the two time periods based on both MMC+ and AUROC behavior.

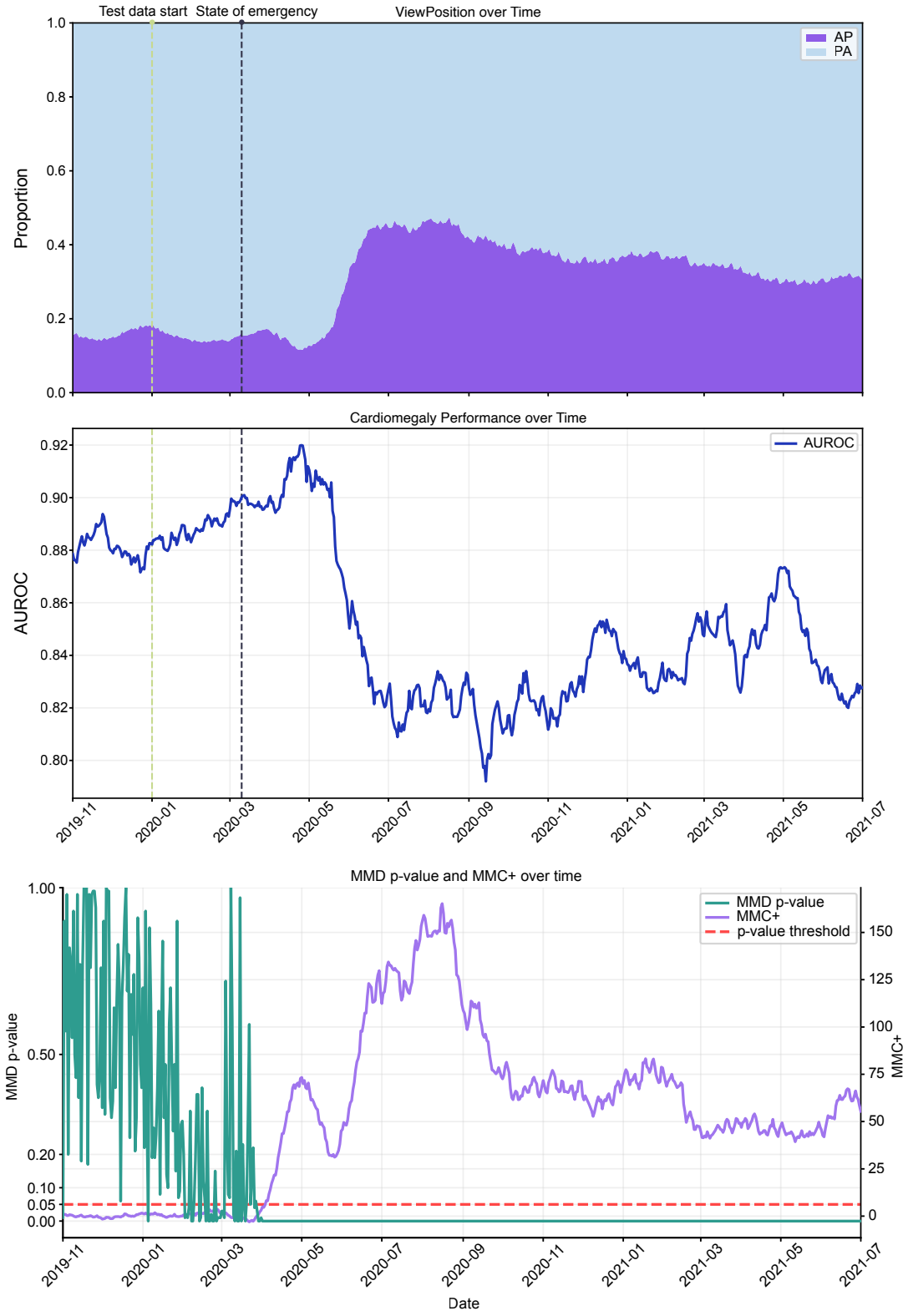


Figure 5. (top) Comparison of View Position distribution and cardiomegaly performance over time as measured by AUROC. The top plot shows the distribution of View Position (AP and PA), and the bottom shows cardiomegaly performance. In both, the dashed vertical line marks March 10, 2020 (bottom) Comparison of MMD-based p-value and MMC+ over time. The light-blue line represents the MMD p-value, while the purple line tracks the MMC+ values over the same period. The red dashed line indicates the *p*-value threshold (set at 0.05).

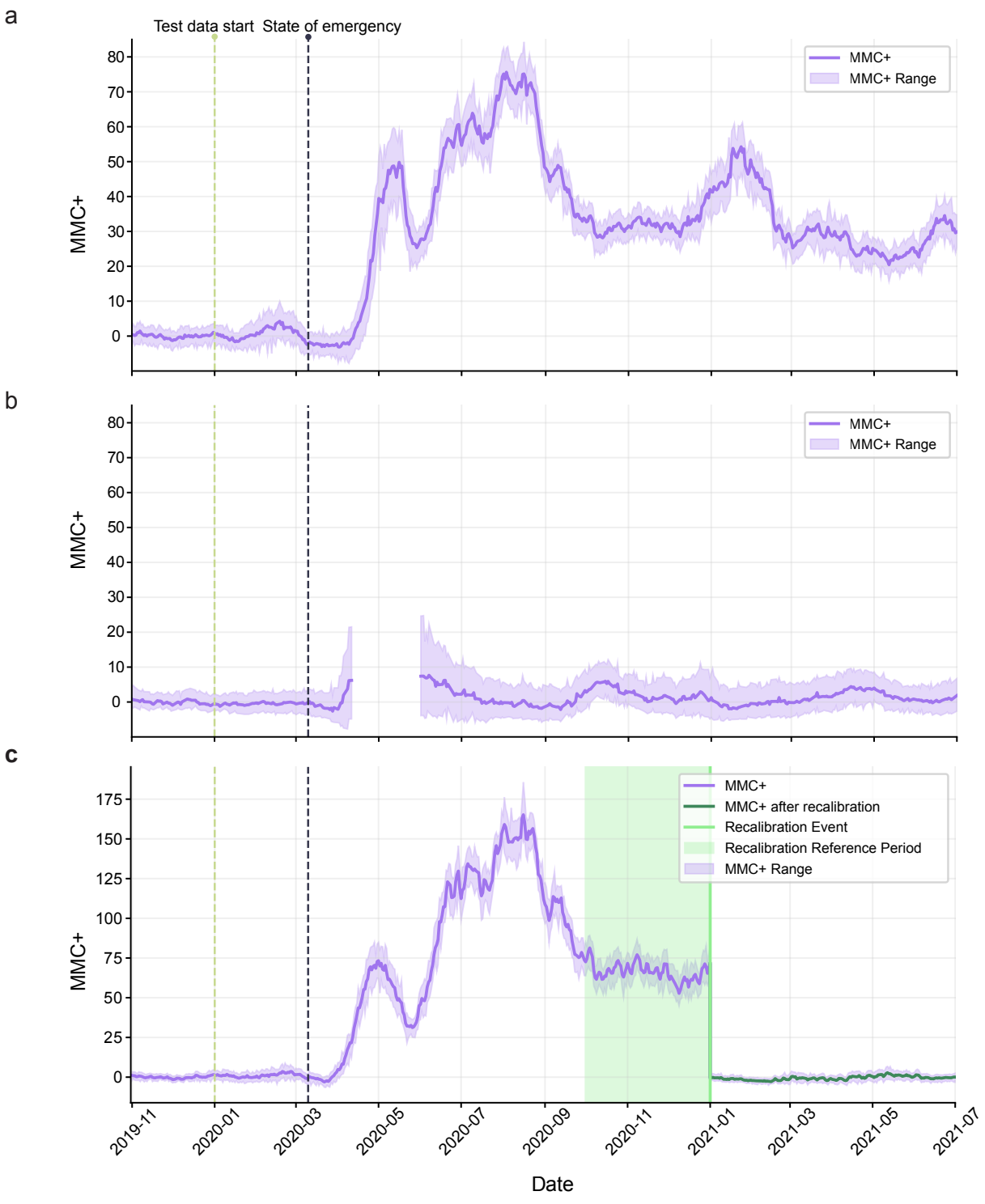


Figure 6. a, b: Plots that show the evolution of MMC+ over time for two different points of care: (a) the ER Point of Care and (b) the ambulatory center operating room Point of Care. In both, the purple line represents the weighted MMC+ values, while the purple shaded area indicates the range or variability around the MMC+. **c**: Plot of MMC+ over time with a re-calibration event. MMC+ recalibration on January 1st, 2021, using a new reference window from October 1st, 2020, to December 31st, 2020. The purple line shows MMC+ values before recalibration, and the dark green line indicates MMC+ values after recalibration. The purple shaded area represents the MMC+ range or variability. The recalibration event is marked by the vertical line, with the recalibration reference period highlighted in green.

3 Discussion

Our work addresses the pressing need for continuous monitoring in medical imaging AI applications. The rapid integration of AI into imaging demands monitoring methods that extend beyond initial validation stages. Ongoing, interpretable monitoring is essential to promptly detect issues such as model drift, feature drift, and shifts in input data. Current practices, which often rely on infrequent reviews and user feedback, are insufficient. They are too slow to respond to emerging problems and lack the depth required to understand the nuances of model behavior. Real-time performance monitoring, while ideal, is frequently cost-prohibitive and may not effectively identify the underlying causes of performance changes.

The need for adaptive monitoring was particularly evident during the COVID-19 pandemic, which brought rapid changes to healthcare workflows and emphasized the necessity of robust monitoring systems. As clinical protocols evolved and healthcare staff faced increased workloads, automated monitoring became crucial in identifying subtle shifts and ensuring the continued effectiveness of medical imaging AI. However, beyond COVID-19, serial chest radiographs are common in hospitalized patients, and often have vast variations in acquisition techniques (projection - posterior-anterior versus anterior-posterior; portable versus non-portable acquisition devices and positions) and patient condition and abnormalities (variations in lung volumes, upright versus supine, rotation, support devices and catheters, and extent and type of abnormalities). These variations make interpretation of radiographs subjective and error-prone for radiologists as well as challenge any AI tools. By providing real-time insights and easing the burden on medical professionals and administrators, these systems will play a vital role in navigating workload challenges while maintaining quality care.

Existing monitoring tools are not well-suited for imaging data due to the complexity of image representations. Conventional methods require extensive site-specific training and are challenging to scale across multiple locations, hindering widespread adoption. To overcome these limitations, we developed MMC+ as a modular framework architecture that identifies core requirements for practical drift detection: aggregating information from multiple heterogeneous data sources, quantifying distribution shifts with stable and interpretable measures, incorporating uncertainty quantification to handle variable sample sizes, and leveraging generalizable image representations that eliminate site-specific training. By establishing architectural principles rather than prescribing specific implementations, MMC+ can be adapted to different clinical contexts while maintaining core monitoring capabilities.

Our framework emphasizes essential properties that effective drift monitoring components should possess. Distribution comparison methods should provide stable quantification that decouples effect size from sample size, avoiding over-sensitivity to statistically significant but practically irrelevant shifts. Components should be robust across varying sample sizes and handle practical challenges such as evolving data characteristics, making them reliable for real-world clinical environments where protocols and populations change over time.

As previous works have noted[21], our findings demonstrate that data drift does not invariably correlate directly to performance degradation; in some cases, performance may even improve. This underscores the importance of monitoring both the model outputs and the input data to gain a comprehensive understanding of the system's behavior. We observed that when MMC+ values remain within acceptable limits, the model's performance is stable. However, significant deviations in MMC+ warrant performance audits, model performance becomes unpredictable. We also observer that monitoring different operational sites (such as ER vs PostOp) separately, when feasible, can enhance detection accuracy and provide more granular insights into the model's performance across diverse datasets. We observed a clear connection between MMC+ values and model performance. Stable MMC+ values correspond to consistent performance levels, whereas large deviations in MMC+ result in variable and unpredictable performance. This insight is crucial as it offers a cost-effective compromise between continuous performance monitoring and periodic assessments.

In practical terms, hospitals can deploy MMC+ for continuous, real-time monitoring of AI models. This assists clinicians in identifying when AI support may be unreliable or require attention, thereby enhancing patient safety. However, obstacles remain. Existing methods might miss subtle yet important declines in model performance, and relying solely on proxies without verifying actual performance could compromise patient outcomes.

The wider impact and applicability of MMC+ are significant. The framework architecture can be adapted across different medical imaging modalities, stressing the importance of ongoing AI monitoring after deployment to ensure reliability. Incorporating uncertainty bounds aids in managing fluctuations in number of imaging studies over time and avoiding spurious drift measurements when data sizes are small. Transitioning to generalizable representations, such as those from foundation models like MedImageInsight reduces retraining needs, enhancing scalability and ease of deployment. This approach simplifies processes and reducing costs. Temporal assessment of abnormalities in context of changing acquisition techniques and radiographic quality along with evolving disease severity and developing and/or resolving coexisting comorbidities, is part of routine clinical interpretation of chest radiographs. Their impact on AI performance and the need for monitoring are thus of paramount importance. The highly subjective nature of projectional radiography and substantial inter-observer variations among radiologists underscore the importance of an automated and efficient real-time monitoring such as the MMC+. Such human observer variations often necessitate multi-reader participation which are both expensive and inefficient.

Our results highlight the importance of choosing drift quantification methods that separate statistical significance from practical relevance, particularly when large sample sizes can amplify inconsequential shifts. Methods that conflate these concerns often lead to volatile and difficult-to-interpret signals. In contrast, approaches that directly quantify distributional dissimilarity offer more stable and intuitive representations by measuring meaningful changes rather than significance levels. These findings suggest that measuring effect size rather than testing hypotheses presents a more effective foundation for tracking and visualizing drift in real-world applications.

We acknowledge that while this study focused on chest X-ray data, the MMC+ framework was designed with flexibility and generalizability as core principles. The modular architecture enables adaptation to diverse imaging modalities through appropriate choice of distance metrics and representations. Although we have not validated the framework on three-dimensional imaging data such as CT or MRI scans, the design accommodates different data types. While we demonstrate foundation model embeddings from MedImageInsight, which has been validated across CT, MRI, mammography, and text, the framework can incorporate alternative representation methods suited to specific modalities or computational constraints. Future work should evaluate the framework across additional medical imaging tasks to characterize how different component choices affect performance, and address operational considerations such as optimal alert thresholds, resource-limited environments, and regulatory requirements across healthcare systems.

Another limitation of our study is the lack of validation in a broader set of common/contemporary clinical indications with substantially reduced COVID-19 prevalence. However, given the non-specific nature of COVID-19 related findings on chest radiographs, and the scope of abnormalities from lungs to the bones assessed in our study, we believe that our results will be generalizable across other clinical indications.

MMC+ provides a scalable framework architecture for drift monitoring in medical imaging AI that accommodates appropriate methods for distribution comparison and leverages generalizable input representations. The framework addresses limitations of existing approaches by eliminating site-specific training requirements and enabling adaptation across various settings. Furthermore, the architectural design aligns with FDA guidance on continuous post-market surveillance for AI-enabled medical devices, providing signals when performance may deviate from acceptable bounds while meeting evolving regulatory standards. Future work should evaluate the framework across additional imaging modalities, characterize how component choices affect detection sensitivity, and address operational considerations such as alert thresholds and resource constraints. Overall, MMC+ advances dependable and scalable AI monitoring in healthcare by establishing principles for robust drift detection systems.

4 Materials and Methods

4.1 Metric Development and Calculation

In our previous work, we introduced the CheXstray framework, which laid the foundation for generating the multi-modal concordance (MMC) metric. In outline, MMC is an aggregate of distance measurements across multiple data elements including image metadata, image appearance, and model predictions, providing a single value representing the overall dissimilarity between two sets of medical exams. For each component data element, a unimodal distance metric is calculated using appropriate statistical methods. These distance metrics are then combined through a process of normalization, weighting, and aggregation to produce a unified MMC score.

In this work, we introduce an updated MMC+ framework and with it we have identified essential properties for effective distribution comparison, introduced uncertainty quantification methods to properly reflect sampling variability, and demonstrated how foundation model embeddings satisfy requirements for generalization across sites without site-specific training. We show how components satisfying these properties address practical deployment challenges while maintaining the original aggregation and normalization approach. In the following sections, we detail these design principles and demonstrate their effectiveness through specific instantiations.

The multi-modal concordance (MMC+) metric quantifies the degree of drift or change between a *reference window*, R, representing a dataset with well-characterized model performance, and a *detection window*, denoted by W, with unknown model behavior. The reference window is typically collected shortly before a model is deployed, using data with known ground truth to assess model performance. The detection window, on the other hand, consists of the set of samples within a sliding temporal window of fixed duration used to assess drift over time. If the MMC+ indicates a high level of dissimilarity between the detection window and the reference window, it suggests that the model is currently operating on data upon which its performance is not well characterized.

The detection window is defined by two key temporal parameters: duration and stride. The duration, τ, specifies the total length of time that each window covers, while the stride, δ, determines the time interval between the start of one window and the start of the next. As such, if a data stream consists of items x_i (each representing both an image and its associated metadata)

arriving at time t_i, the detection window is constructed according to:

$$\begin{aligned} W_k &= \{x_i \mid t_k - \tau < t_i \leq t_k\} \\ t_{k+1} &= t_k + \delta \end{aligned} \tag{1}$$

Each detection window W_k serves as the basis for which the MMC+ metric is computed. As the detection window slides over time, new windows are generated at each step, enabling continuous monitoring of model performance. The MMC+ metric quantifies the drift between these detection windows and the reference window, providing a robust measure of how well the model is performing on current data relative to the established baseline. By systematically calculating MMC+ to each detection window, we can assess and track the stability of the data stream over time. For clarity, we drop the notation indicating the time at which the window was taken, as the metric calculation is generic with respect to time and does not depend on when the window was collected.

Distribution comparison is the core component for quantifying differences between the reference and detection windows. The original MMC metric relied on statistical hypothesis tests for this purpose. However, further experimentation revealed limitations in this approach, prompting us to establish principles for effective drift quantification.

In practice, distribution comparison methods should satisfy several essential requirements. First, they should provide well-defined measures even when categorical values appear in one set but not the other, a common occurrence when new equipment or protocols are introduced. Second, they should quantify distributional dissimilarity in a way that is robust to sample size variations, separating the magnitude of shifts from statistical significance. Third, the methods should measure distributional differences directly rather than conflating distance with uncertainty estimation. For continuous variables, methods should handle multivariate data and detect shifts in joint distributions, capturing both changes in probability density and shifts in distribution location or shape.

The MMC+ framework accommodates any comparison methods that satisfy these requirements. For this implementation, we employ non-parametric probability distribution comparison measures appropriate for both discrete categorical variables and continuous multivariate data, avoiding distributional assumptions while providing stable, interpretable quantification of distributional dissimilarity. Specific methods and their mathematical formulations are described a subsequent section 4.4 with additional details in the Appendix.

Effective drift monitoring requires quantifying not only distributional differences but also the uncertainty in those measurements, particularly when detection window sample sizes vary over time. The original CheXstray framework addressed sample size sensitivity through bootstrap oversampling to a fixed size, but this approach discards information about measurement uncertainty at low sample sizes.

MMC+ adopts an alternative approach that explicitly quantifies uncertainty in distribution comparison measures. Following the Flapjack algorithm[24], we repeatedly subsample from the reference window to estimate both the central tendency and variability of the comparison measure. Specifically, we draw two independent random subsets $U, V \subseteq R$ without replacement, where subset sizes match the detection window size: $|U| = |V| = |W|$. This ensures fair comparisons unbiased by differing sample sizes. The procedure requires $|W| < |R|$ for all detection windows, necessitating a sufficiently large reference window in practice. We repeat this sampling N_s times to generate multiple subsets $\{U_n\}_{n=1}^{N_s}$ and $\{V_n\}_{n=1}^{N_s}$, which are used to compute a distribution of comparison measure estimates:

$$D(W,R) = \{d_n(W,R)\}_{n=1}^{N_s} = \{\mathscr{D}(U_n,W) - \mathscr{D}(U_n,V_n)\}_{n=1}^{N_s} \tag{2}$$

where $\mathscr{D}$ denotes the appropriate comparison measure for each variable type. The baseline correction $\mathscr{D}(U_n,V_n)$ accounts for inherent variability within the reference window due to finite sampling, isolating the true distributional difference between W and R beyond what sampling variation alone would produce. The number of iterations N_s can be adjusted to balance computational cost against precision. For notational simplicity, we subsequently denote this set as D, with the dependence on R and W implicit.

We report the mean value of D over samples as the estimated distance and the minimum and maximum values across D to obtain an estimate of the uncertainty of the distance metric. Since MMC+ aggregates comparison measures from multiple components, we apply this sampling procedure to each component independently, computing distance estimates using the appropriate measure for each data type. We denote component-specific measures with superscript c, where $d_n^{(c)}$ represents the distance estimate for the c^{th} component from the n^{th} subsample:

$$\psi^{(c)} = \frac{1}{N_s}\sum_{n=1}^{N_s} d_n^{(c)}, \quad \langle \psi_U^{(c)}, \psi_L^{(c)} \rangle = \langle \max_n d_n^{(c)}, \min_n d_n^{(c)} \rangle \tag{3}$$

where $\psi^{(c)}$ provides the central estimate of distributional dissimilarity for component c, while $\psi_U^{(c)}$ and $\psi_L^{(c)}$ capture the range of estimates across all N_s subsamples, quantifying uncertainty in the measurement.

The MMC+ metric combines distance metrics from all components c into a single metric. Since the metrics for each component have different ranges, we first normalize them. For the c-th component, we calculate normalization parameters by finding the mean $\zeta^{(c)}$ and standard deviation $\eta^{(c)}$ of the comparison measure $\psi^{(c)}$ between each possible detection window W_k within the reference set R and the remainder of the reference set $\hat{R}_k$, defined as:

$$\hat{R}_k = \{x_i \mid x_i \in R,\ x_i \notin W_k\} \tag{4}$$

In addition to normalization, we weight each component to reflect its importance or reliability in the final unified score. For component c, we calculate a weight value $\alpha^{(c)}$ based on the absolute value of the Pearson correlation coefficient (ρ) between the model performance and the distributional comparison measure. Here, performance is measured using the area under the receiver operating characteristic curve (AUROC), denoted as $a(W_k)$, where W_k represents each detection window within the reference set R. Like the normalization calculation, the comparison measure is computed between W_k and the remainder of the reference set, $\hat{R}_k$. The weight is computed as:

$$\alpha^{(c)} = \left| \rho \left(\left\{ \psi^{(c)}(\hat{R}_k, W_k) \right\}_k, \{a(W_k)\}_k \right) \right| \tag{5}$$

Combining the normalized and weighted measures from C components, we calculate the multimodal concordance metric (MMC+) for the detection window W as:

$$\text{MMC+} = \sum_{c=1}^{C} \alpha^{(c)} \frac{\psi^{(c)} - \zeta^{(c)}}{\eta^{(c)}} \tag{6}$$

To account for uncertainty, we also compute the upper and lower bounds of MMC+ denoted as

$$\langle \text{MMC+}_U,\ \text{MMC+}_L \rangle = \sum_{c=1}^{C} \alpha^{(c)} \frac{\left\langle \psi_U^{(c)},\ \psi_L^{(c)} \right\rangle - \zeta^{(c)}}{\eta^{(c)}} \tag{7}$$

Here, MMC+$_U$ and MMC+$_L$ are the upper and lower bounds of the distance metric for component c, obtained from the uncertainty estimates as described in the previous section. By weighting each component according to its correlation with model performance, we emphasize metrics that are more predictive of performance changes, enhancing the sensitivity and reliability of MMC+. The upper and lower bounds provide a range for MMC+, offering insight into potential variability due to sampling or measurement uncertainties.

4.2 Dataset

Our dataset was created by querying an institutional radiology database at Massachusetts General Hospital for all imaging studies matching the following criteria: 1) modality of "XR" (X-ray), 2) an institution-specific examination code with one of two values corresponding to "X-Ray Chest 1 View" and "X-Ray Chest PA And Lateral 2 Views", 3) a study date from within the two-year period between 2019-07-01 and 2021-07-01 ("yyyy-mm-dd" format, inclusive), and 4) patient age of 18 years or older. This dataset was chosen to reflect the full set of studies that would likely be routed to an X-ray analysis model if it were deployed in clinical practice at Massachusetts General Hospital, as well as to capture the period of time before, during, and after the initial wave of the COVID-19 pandemic.

This process yielded 90,713 studies of which 90,581 (from 60,062 distinct patients) were successfully retrieved from the institutional Picture Archive and Communication System (PACS) along with their radiology reports. To create ground truth labels for this dataset, we used a large language model (QWEN 1.5-72B) to extract labels for each of a set of nine findings by few-shot prompting the model with a small number of positive and negative examples. We characterized the accuracy of this process on a subset of this same dataset in previously published work[25], and found it to be highly accurate for most findings. Following the Chexpert convention[23], the nine selected findings were: atelectasis, cardiomegaly, edema, lung opacity, pleural effusion, pleural other, pneumonia, pneumothorax, and support devices. We excluded the following findings due to inconsistency when extracting ground truth labels: consolidation, enlarged cardiomediastinum, fracture, lung lesion, and "no finding". We note that ground truth labels were required to assess the relationship of drift metrics to performance, but would not be required to use the MMC+ framework to monitor for drift in practice.

We divided the dataset into three subsets based on the date of acquisition as follows: studies acquired in the first three months (July 2019 through September 2019) were used for model training and validation, studies from the next three months (October 2019 through December 2019) were used as reference set for the drift metric, and the remaining studies covering an 18-month period from the start of January 2020 until the end of June 2021 were used for the drift experiments. This division was chosen to approximately reflect a model development process. In such a process a model would first be developed on retrospective images, then validated on recent images before being deployed prospectively on new images. Characteristics of the three dataset partitions are presented in Table 2, and the distribution of findings across partitions is presented in Table A.2.

	Training	Reference	Test
Start Date	2019-07-01	2019-10-01	2020-01-01
End Date	2020-09-31	2020-12-31	2021-07-01
Total Images	12,478	14,936	78,181
Age (years, mean $\pm$ std)	60.2 $\pm$ 18.48	59.53 $\pm$ 18.22	58.64 $\pm$ 17.87
Female (%)	48.12	49.44	46.03
Inpatient (%)	14.74	14.44	27.84
Emergency (%)	31.53	29.37	27.68
Portable (%)	28.15	27.79	51.38

Table 2. Summary of dataset characteristic statistics.

4.3 Classification Model

For the purposes of our experiments, the AI model whose performance we monitored was a classification model trained to detect multiple findings in a chest radiograph. This model was based on the DenseNet-121 architecture, pretrained on ImageNet-1k dataset[26,27]. The output layer was adapted to consist of nine elements, one for each of the nine findings, followed by a sigmoid activation to yield nine independent (multi-label) predictions. The model was trained on the images from the training set, defined in Section 4.2. Training was conducted with a batch size of 64, a learning rate of 1×10^{-4}, a binary cross-entropy loss function, the Adam optimizer and a ReduceLROnPlateau learning rate schedule. The model was trained exclusively on frontal (AP and PA) images (lateral views were excluded) across all points of care. Training was performed using a single A100-40GB GPU (NVIDIA, Santa Clara, CA, USA).

4.4 MMC+ Experimental Implementation

The preceding sections describe the architectural principles and requirements of the MMC+ framework. This section details how we instantiate the framework for our chest X-ray experiments, specifying the concrete methods and data elements used to satisfy those requirements.

For distribution comparison of continuous variables, we employ the Wasserstein distance, which effectively captures differences in multivariate distributions and handles shifts in both location and shape. Our implementation of the Wasserstein distance uses the Python Optimal Transport (POT) library[28]. For discrete categorical variables, we use the Hellinger distance, which provides well-defined measures even when categories differ between sets and remains stable across varying sample sizes. Mathematical formulations for these distance metrics are provided in the Appendix.

MMC+ aggregates information from three types of data elements: image metadata, image appearance, and model predictions. While the framework is general and can accommodate various measurements, we describe here our specific instantiation for chest X-ray monitoring.

Every radiological image acquired in clinical practice is stored within a standardized DICOM (Digital Imaging and Communications in Medicine) file alongside various metadata attributes that describe acquisition parameters of the image, patient characteristics, and the clinical context of the acquisition. For our experiments, we selected a number of DICOM attributes that may affect model performance on chest radiographs and used each as an individual element within MMC+. A small number of additional attributes were sourced from the institutional Radiology Information System (RIS). Table A.1 in the appendix lists the relevant attributes. Each metadata attribute is treated as either a discrete or continuous variable and compared using the appropriate distance metric (Hellinger or Wasserstein).

In our previous implementation, we used a 128-dimensional latent vector produced by a VAE to represent the appearance of medical images. The VAE was trained on chest radiographs from the PadChest dataset[29], and the resulting latent vector served as the basis for calculating distance metrics to monitor shifts in image data. While this approach provided a functional, low-dimensional representation, it posed several challenges. Training the VAE required significant computational resources and data preparation which when in operation would need to be repeated at each site. Moreover, the VAE's limited expressiveness made it insufficient for capturing the complexity of medical images, particularly across imaging modalities.

To overcome these limitations, we transitioned to using MedImageInsight[22], a foundation model designed specifically for medical image analysis, which provides a 1024-dimensional embedding. Unlike the VAE, MedImageInsight does not require site-specific training, offering a powerful pre-trained solution that can be applied to a wide range of medical images out of the box. The larger 1024-dimensional embedding is significantly more expressive, capturing a richer and more detailed representation of image appearance, enabling the system to better generalize across various clinical settings and imaging modalities.

By leveraging MedImageInsight's pre-trained, high-dimensional embeddings, we eliminate the need for extensive, site-

specific training processes, simplifying deployment and reducing the operational complexity of maintaining the system across different locations. MedImageInsight's generalizability allows it to adapt seamlessly to new imaging environments and protocols without requiring retraining, which not only enhances the robustness and accuracy of our image appearance monitoring framework but also ensures the system is more scalable and easier to maintain. This transition ultimately reduces the operational burden, making the system more efficient in real-world clinical environments. We compute the Wasserstein distance on these 1024-dimensional embeddings to quantify appearance drift.

The raw classifier output (after the sigmoid activations) consists of 9 independent values in the range 0–1, one for each of the findings. These values carry information about both the likely class distribution of the datastream and the model's confidence in its predictions. This nine element vector is used directly as a multidimensional variable, with the Wasserstein distance computed on this continuous multivariate representation to quantify drift in model predictions.

Author Contributions

J.M., C.P.B., M.P.L., I.T. and A.E. conceived the study. J.M. F.J.D., X.Y., and C.P.B. designed and conducted experimental studies. G.D. and M.K. prepared data and contributed to the study design. J.M., F.J.D, and C.P.B. wrote the main manuscript text. All authors reviewed the manuscript.

Competing Interests

This work was partially supported by Microsoft, a global technology company that offers both paid and free products, including AI monitoring solutions which are potentially associated with patents.

Funding Statement

This project did not receive specific funding.

Data Availability

This manuscript used a large number of chest radiographs from Massachusetts General Hospital. This data is not publicly available due to patient privacy restrictions.

Code Availability

Source code is available at https://github.com/FJDorfner/MedImaging-ModelDriftMonitoring/tree/MMC_plus.

References

1. van Leeuwen, K. G., Schalekamp, S., Rutten, M. J., van Ginneken, B. & de Rooij, M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur. radiology* **31**, 3797–3804 (2021).
2. West, E., Mutasa, S., Zhu, Z. & Ha, R. Global trend in artificial intelligence–based publications in radiology from 2000 to 2018. *Am. J. Roentgenol.* **213**, 1204–1206 (2019).
3. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine* **3**, 1–8 (2020).
4. Mehrizi, M. H. R., van Ooijen, P. & Homan, M. Applications of artificial intelligence (ai) in diagnostic radiology: a technography study. *Eur. radiology* **31**, 1805–1811 (2021).
5. Tadavarthi, Y. *et al.* The state of radiology ai: considerations for purchase decisions and current market offerings. *Radiol. Artif. Intell.* **2**, e200004 (2020).
6. Park, S. H. *et al.* Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* **306**, 20–31 (2023).
7. Omoumi, P. *et al.* To buy or not to buy—evaluating commercial ai solutions in radiology (the eclair guidelines). *Eur. radiology* **31**, 3786–3796 (2021).
8. Feng, J. *et al.* Clinical artificial intelligence quality improvement: towards continual monitoring and updating of ai algorithms in healthcare. *npj Digit. Medicine* **5**, 66, DOI: 10.1038/s41746-022-00611-y (2022).

9. Rajpurkar, P. & Lungren, M. P. The current and future state of ai interpretation of medical images. *New Engl. J. Medicine* **388**, 1981–1990, DOI: 10.1056/NEJMra2301725 (2023).

10. Pianykh, O. S. *et al.* Continuous learning ai in radiology: implementation principles and early applications. *Radiology* **297**, 6–14 (2020).

11. Lacson, R., Eskian, M., Licaros, A., Kapoor, N. & Khorasani, R. Machine learning model drift: Predicting diagnostic imaging follow-up as a case example. *J. Am. Coll. Radiol.* **19**, 1162–1169 (2022).

12. Rahmani, K. *et al.* Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. *Int. J. Med. Informatics* 104930, DOI: 10.1016/j.ijmedinf.2022.104930 (2023).

13. Ross, C. Ai gone astray: how subtle shifts in patient data send popular algorithms reeling, undermining patient safety. *STAT* (2022).

14. U.S. Food and Drug Administration. Artificial intelligence–enabled device software functions: Lifecycle management and marketing. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing (2023). FDA Guidance Document.

15. Daye, D. *et al.* Implementation of clinical artificial intelligence in radiology: who decides and how? *Radiology* **305**, 555–563 (2022).

16. Bizzo, B. C. *et al.* Addressing the challenges of implementing artificial intelligence tools in clinical practice: principles from experience. *J. Am. Coll. Radiol.* **20**, 352–360 (2023).

17. Subasri, V. *et al.* Cyclops: a unified framework for data extraction and rigorous evaluation of ml models for clinical use-cases. *Mach. Learn. for Healthc. 2022–Clinical Abstr. Software, Demo Track* (2022).

18. Duckworth, C. *et al.* Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during covid-19. *Sci. reports* **11**, 23017 (2021).

19. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

20. Merkow, J. *et al.* CheXstray: A Real-Time Multi-Modal Monitoring Workflow for Medical Imaging AI. In Greenspan, H. *et al.* (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, vol. 14222, 326–336, DOI: 10.1007/978-3-031-43898-1_32 (Springer Nature Switzerland, Cham, 2023). Series Title: Lecture Notes in Computer Science.

21. Kore, A. *et al.* Empirical data drift detection experiments on real-world medical imaging data. *Nat. Commun.* **15**, 1887, DOI: 10.1038/s41467-024-46142-w (2024).

22. Codella, N. C. F. *et al.* Medimageinsight: An open-source embedding model for general domain medical imaging (2024). 2410.06542.

23. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence* (2019).

24. Papp, T. & Sherlock, C. Bounds on wasserstein distances between continuous distributions using independent samples. *arXiv preprint arXiv:2203.11627* (2022).

25. Dorfner, F. J. *et al.* Comparing Commercial and Open-Source Large Language Models for Labeling Chest Radiograph Reports. *Radiology* DOI: 10.1148/radiol.241139 (2024). Publisher: Radiological Society of North America.

26. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks (2018). 1608.06993.

27. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255, DOI: 10.1109/CVPR.2009.5206848 (2009).

28. Flamary, R. *et al.* Pot: Python optimal transport. *J. Mach. Learn. Res.* **22**, 1–8 (2021).

29. Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Med. Image Analysis* **66**, 101797, DOI: 10.1016/j.media.2020.101797 (2020).

30. Oosterhoff, J. & van Zwet, W. R. *A note on contiguity and Hellinger distance* (Springer, 2012).

31. Panaretos, V. M. & Zemel, Y. Statistical aspects of wasserstein distances. *Annu. review statistics its application* **6**, 405–431 (2019).

A Appendix

Tag	Keyword	Type	Description
0008,0060	Modality	Discrete	DX "Digital Radiography" or CR "Computed Radiography".
0008,0070	Manufacturer	Discrete	Manufacturer of the scanner that acquired the image.
0008,1090	ManufacturerModelName	Discrete	Model name of the scanner that acquired the image.
0010,0040	Patient Sex	Discrete	Sex of the patient.
0018,0060	KVP	Continuous	Kilo-volt peak of the source X-Ray tube.
0018,1151	XRayTubeCurrent	Continuous	Current of the source X-Ray tube.
0018,1153	ExposureInμAs	Continuous	Radiation exposure in μAs.
0018,5100	ViewPosition	Discrete	Radiographic view (e.g. posterior/anterior, lateral).
0028,0004	PhotometricInterpretation	Discrete	Interpretation of pixel data.
0028,0010	Rows	Continuous	Number of rows in the image.
0028,0011	Columns	Continuous	Number of columns in the image.
0028,0030	PixelSpacing	Continuous	Physical spacing between pixels.
0028,0101	BitsStored	Discrete	Pixel bit-depth.
(RIS)	Patient Age	Continuous	Age of the patient in years.
(RIS)	Point of Care	Discrete	Hospital location.
(RIS)	Is Stat	Discrete	Whether image was marked for urgent review.
(RIS)	Exam Code	Discrete	Internal procedure code (see Section 4.2).

Table A.1. Table of DICOM and RIS attributes used for experiments including the tags, keywords, types (discrete or continuous), and descriptions of the data fields utilized in the analysis. These attributes include both imaging metadata (e.g., modality, exposure, pixel data) and patient-specific information (e.g., age, sex, point of care).

	Training	Validation	Test
Total Images	12,476	14,931	78,168
Atelectasis	3,610 (28.94%)	4,204 (28.16%)	25,966 (33.22%)
Cardiomegaly	1,838 (14.73%)	2,048 (13.72%)	10,381 (13.28%)
Edema	1,455 (11.66%)	1,582 (10.60%)	12,022 (15.38%)
Lung Opacity	5,554 (44.52%)	6,479 (43.39%)	38,252 (48.94%)
Pleural Other	637 (5.11%)	717 (4.80%)	3,810 (4.87%)
Pleural Effusion	2,239 (17.95%)	2,491 (16.68%)	16,742 (21.42%)
Pneumonia	1,668 (13.37%)	2,061 (13.80%)	13,194 (16.88%)
Pneumothorax	287 (2.30%)	350 (2.34%)	2,548 (3.26%)
Support Devices	3,190 (25.57%)	3,645 (24.41%)	25,694 (32.87%)

Table A.2. Distribution of findings across training, validation, and test datasets. Values represent the number of positive cases with prevalence percentages in parentheses.

A.1 Distance Metric Formulations

This section provides the mathematical formulations for the distance metrics used in our implementation of MMC+.

For discrete variables (such as scanner model name or photometric interpretation) with K discrete values, we calculate distances between two generic sets of data points, $P = \{p_i\}_{i=1}^{N}$ and $Q = \{q_i\}_{i=1}^{M}$ where $p_i, q_i \in \{1, 2, \ldots, K\}$, using the Hellinger distance[30], which is defined as

$$\mathscr{H}(P,Q) = \sqrt{\frac{1}{2}\sum_{k=1}^{K}\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(p_i = k)} - \sqrt{\frac{1}{M}\sum_{i=1}^{M}\mathbb{I}(q_i = k)}\right)^2}, \tag{8}$$

where $\mathbb{I}(\cdot) \in \{0,1\}$ is the indicator function.

For d-dimensional continuous variables, $p_i, q_j \in \mathbb{R}^d$, we use the Wasserstein distance (otherwise known as the earth-mover's distance). In its general form, the Wasserstein distance between two distributions $p(x)$ and $q(x)$ is given by the cost of the

optimal transport plan that moves the probability mass of $p(x)$ to $q(x)$, where the cost is defined as equal to the amount of probability mass moved multiplied by distance moved in the underlying space[31]. For empirical distributions, where the distribution is formed of point masses at sample points $P = \{p_i\}_{i=1}^{N}$ and $Q = \{q_j\}_{j=1}^{M}$, the optimal transport plan, and therefore Wasserstein distance, may be found by computing Euclidean distances between all pairings of points in P and Q, and solving the following linear program:

$$\begin{aligned} \mathscr{W}(P,Q) \quad = \quad & \min_{\gamma} \sum_{i=1}^{N}\sum_{j=1}^{M} \gamma_{i,j}\|p_i - q_j\|_2 \\ & \text{s.t.} \quad \sum_{i=1}^{N} \gamma_{i,j} = \frac{1}{M}, \quad j = 1,\ldots,M \\ & \qquad \sum_{j=1}^{M} \gamma_{i,j} = \frac{1}{N}, \quad i = 1,\ldots,N \\ & \qquad 0 \leq \gamma_{i,j} \leq 1, \quad i = 1,\ldots,N, j = 1,\ldots,M \end{aligned} \tag{9}$$

where $\gamma_{i,j} \in \mathbb{R}$ is the quantity of probability mass moved from initial point p_i in the first distribution to end point q_j in the second distribution. In our implementation, we use the open-source Python Optimal Transport (POT) library[28] to calculate the solution to this optimal transport problem.