ClinValAI: A framework for developing Cloud-based infrastructures for the External Clinical Validation of AI in Medical Imaging

Ojas A. Ramwala[1], Kathryn P. Lowry[2], Daniel S. Hippe[3], Matthew P.N. Unrath[4], Matthew J. Nyflot[2,5], Sean D. Mooney[6], Christoph I. Lee[2,7,8†]

[1]*Department of Biomedical Informatics and Medical Education, University of Washington School of Medicine, Seattle, Washington, 98195, USA*
[2]*Department of Radiology, University of Washington School of Medicine, Seattle, Washington, 98195, USA*
[3]*Clinical Research Division, Fred Hutchinson Cancer Center, Seattle, WA, 98109, USA*
[4]*Pariveda Solutions, Seattle, Washington, 98101, USA*
[5]*Department of Radiation Oncology, University of Washington School of Medicine, Seattle, 98195, USA*
[6]*Center for Information Technology, National Institutes of Health, Bethesda, Maryland, 20892, USA*
[7]*Department of Health Systems and Population Health, University of Washington School of Public Health, Seattle, Washington, 98195, USA*
[8]*Northwest Screening and Cancer Outcomes Research Enterprise, University of Washington, Seattle, Washington, 98195, USA*
[†]*Email: stophlee@uw.edu*

Artificial Intelligence (AI) algorithms showcase the potential to steer a paradigm shift in clinical medicine, especially medical imaging. Concerns associated with model generalizability and biases necessitate rigorous external validation of AI algorithms prior to their adoption into clinical workflows. To address the barriers associated with patient privacy, intellectual property, and diverse model requirements, we introduce ClinValAI, a framework for establishing robust cloud-based infrastructures to clinically validate AI algorithms in medical imaging. By featuring dedicated workflows for data ingestion, algorithm scoring, and output processing, we propose an easily customizable method to assess AI models and investigate biases. Our novel orchestration mechanism facilitates utilizing the complete potential of the cloud computing environment. ClinValAI's input auditing and standardization mechanisms ensure that inputs consistent with model prerequisites are provided to the algorithm for a streamlined validation. The scoring workflow comprises multiple steps to facilitate consistent inferencing and systematic troubleshooting. The output processing workflow helps identify and analyze samples with missing results and aggregates final outputs for downstream analysis. We demonstrate the usability of our work by evaluating a state-of-the-art breast cancer risk prediction algorithm on a large and diverse dataset of 2D screening mammograms. We perform comprehensive statistical analysis to study model calibration and evaluate performance on important factors, including breast density, age, and race, to identify latent biases. ClinValAI provides a holistic framework to validate medical imaging models and has the potential to advance the development of generalizable AI models in clinical medicine and promote health equity.

Keywords: Artificial Intelligence; Bias; Breast Cancer; Clinical Validation; Cloud Infrastructures; Generalizability; Medical Imaging

1. Introduction

Artificial Intelligence (AI) algorithms have demonstrated encouraging results in the field of biomedical signal[1,2] and image[3–11] processing, electronic health record (EHR) analysis[12], and clinical text processing[13] to provide improved diagnostic outcomes, early intervention strategies, and well-tailored patient-specific management options. The performance of AI algorithms has been on par with radiologists[14] and even better in a few scenarios[15].

However, deep learning models are susceptible to generalizability challenges[16]. Diagnostic AI models have demonstrated deteriorated performance during independent evaluation on datasets reflecting real-world healthcare settings, especially for specific subpopulations[17]. The adoption of such algorithms can have critical implications for patients' safety. Thus, large-scale independent external validation of AI models is imperative before adopting them into clinical workflows.

Nevertheless, there are several barriers to robust evaluation. Since AI vendors are protective of their intellectual property, they may be unwilling to provide their algorithms to health institutions for validation, especially prior to their purchase. Per HIPAA guidelines, medical centers cannot share patient data with commercial organizations without their consent since it contains protected health information. Moreover, different AI algorithms have varying storage and computing requirements. Planning and budgeting for resources to cater to such varying needs can cause substantial financial and cognitive burdens on health systems evaluating multiple AI tools on-premises for clinical adoption. Outsourcing clinical validation work to third-party services can be costly and involve legal and operational complications while sharing access to clinical data.

To address the limited technical guidance on developing methods that can aid in monitoring the performance of AI in clinical medicine[18], we propose ClinValAI – an open-source cloud-agnostic unified framework for establishing robust infrastructures to validate AI algorithms. We customize its functionalities for the clinical validation of AI models for medical imaging applications. Our work aims to enable medical institutions to rigorously evaluate models prior to their integration into clinical workflows. By leveraging our framework, healthcare institutions can screen data from large populations to accurately assess model generalizability and investigate latent biases.

To demonstrate the capabilities of our framework, we used our ClinValAI-based cloud infrastructure to perform large-scale clinical validation of Mirai[19], a state-of-the-art open-source mammography-based AI algorithm for breast cancer risk prediction. We comprehensively evaluate its generalizability on a large and diverse dataset of 26,449 2D screening mammography exams from 14,291 patients, demonstrating the reliability of our work in monitoring AI models and assessing algorithmic bias in healthcare settings. Our framework has the potential to improve a clinical institution's AI model selection process to enhance patient care for their target population.

2. Methods

The clinical validation of AI algorithms can be performed via on-premises as well as cloud-based infrastructures. While medical institutions traditionally trust on-premises setups with their patient data, challenges, including upfront resource investments, scalability issues, and maintenance overhead, can obstruct validation efforts. In contrast, the configurability of cloud-based storage and computational environments, their cost-effective setup and maintenance, built-in network security

and information recovery services, and rapid acquisition render cloud infrastructures an appealing choice for deploying and rigorously validating AI models on large datasets. ClinValAI can be leveraged to establish innovative, effective, and secure cloud-based validation infrastructures. Figure 1 details our conceptual framework for externally validating AI models in clinical medicine.

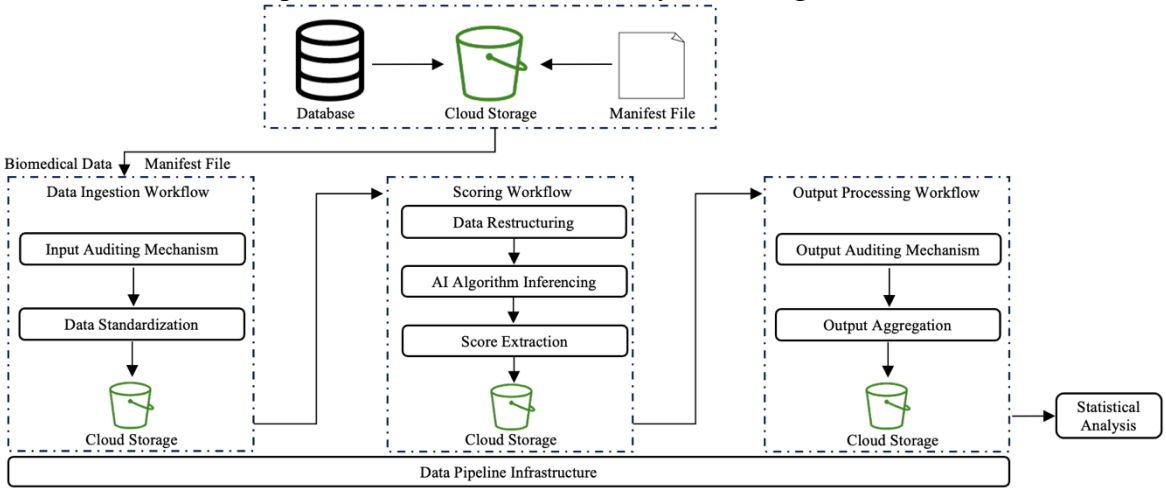

Fig. 1. Conceptual overview of the ClinValAI framework.

2.1. *Preserving Patient Data Privacy*

Patient privacy and information security concerns constrain biomedical data sharing and stymie AI algorithm development and validation efforts[20]. ClinValAI leverages the "Model to Data" (MTD) paradigm[21,22] to validate AI models on private biomedical data. Cloud infrastructure and containerization techniques form the foundation of the MTD framework. Rather than providing direct data access to the vendors, the Dockerized models are uploaded to the cloud host as containers encapsulating the AI algorithms, their dependencies, and other configuration settings required for successfully testing the models on the data stored in the cloud. To address intellectual property concerns, ClinValAI supports license files for Docker images, allowing AI vendors to control access to their AI models. Thus, ClinValAI enables health institutions to preserve patient data within a firewall and run models on medical imaging exams without providing vendors direct data access.

2.2. *Data Pipeline Infrastructure*

ClinValAI features multiple computational pipelines for biomedical data processing and clinical validation of AI algorithms through a combination of series and parallel jobs.

2.2.1. *Workflow Representation*

To comprehensively express the workflow design, we leverage the Workflow Description Language (WDL)[23] due to its comprehensibility and cross-platform interoperability. WDL enables defining pipelines to process and analyze data. WDL necessitates an engine to execute its functionalities. Our proposed framework utilizes miniWDL[24], a WDL execution engine for biomedical applications that functions as a job orchestrator for executing multiple data processing workflows in a parallel

fashion, depending on the available memory and computing resources. The customizability of ClinValAI's workflow representation method bolsters its utilization for the clinical validation of AI.

2.2.2. *Job Scheduling and Batch Processing Orchestration Mechanism*

Our framework is equipped with tools that provision compute instances and communicate with the miniWDL engine and a container job scheduling mechanism to automate infrastructure deployment (Figure 2). It can be further modified for more granular control over those pipelines. After the workflow submission, the WDL script is uploaded to cloud storage, and the job scheduling mechanism is invoked to run a miniWDL container, known as the "head" job container. ClinValAI implements data processing pipelines through the miniWDL engine operating on this container. The head job pulls the WDL script from the cloud storage and, per its instructions, directs the scheduling mechanism to spin up "task" job Docker containers that execute individual components of the workflow. ClinValAI enables the head job containers to spin up multiple sets of task job containers to achieve the parallel execution of computational steps.

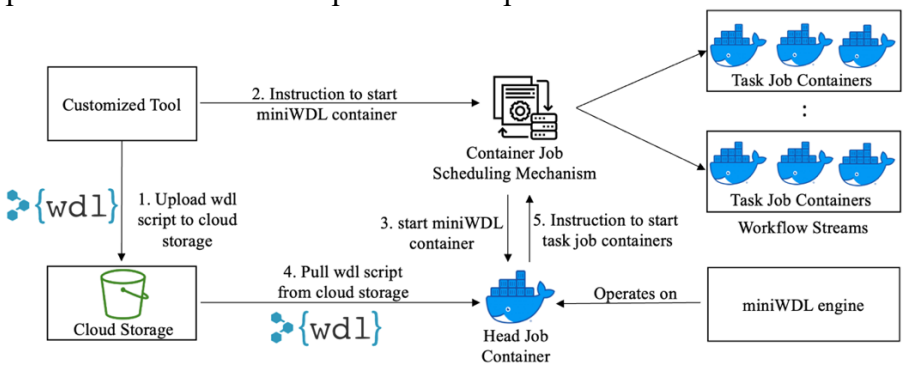

Fig. 2: ClinValAI's job scheduling and batch processing orchestration mechanism.

For our case study on clinical validation of the Mirai algorithm for breast cancer risk prediction on screening mammograms, our ClinValAI-based cloud infrastructure ingests a *set* of compressed files, each representing a *batch* comprising multiple sub-folders corresponding to patients' mammography *exams*. ClinValAI creates multiple execution streams for each set; all exams in one batch are processed serially by leveraging numerous task containers running sequentially. Exams in one batch are scored independently of other batches in a parallel fashion. Thus, ClinValAI's data pipeline enables leveraging the full potential of the cloud computing environment.

In addition to validating AI models using their Docker images, our framework supports customizing Linux Docker images to establish optimized workflows. Rather than dynamically pulling scripts from cloud storage at run-time, ClinValAI facilitates configuring the Docker images at build time. This approach avoids inadvertent version updates in the sequence of instructions during run-time, which could produce inconsistent results. While fetching scripts from cloud storage during run-time is more convenient, baking them into the Docker image enhances reliability.

2.3. *Data Ingestion Workflow*

ClinValAI's data ingestion workflow (Figure 3) is the first of the three stages in the framework. It comprises an input auditing and a data standardization mechanism.

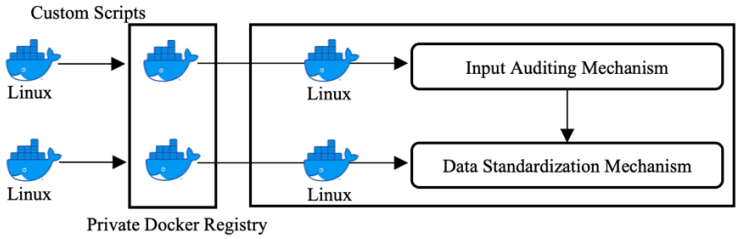

Fig. 3. ClinValAI's data ingestion workflow ensures that inputs are consistent with model prerequisites.

2.3.1. *Input Auditing Mechanism*

ClinValAI's input auditing mechanism performs the vital task of verifying if the data can be processed and are aligned with the model's prerequisites before initiating the scoring process. This can help ensure a sample size that preserves statistical power for meaningful analysis. Through a configured Docker image, it compares the uploaded data with a manifest file and algorithm-specific requirements to verify that the dataset is complete with all the required information.

To validate Mirai, a manifest file comprising the accession numbers, data modality, the corresponding number of images in each exam, image laterality and projection, file sizes, etc., is created. The auditing logic checks for corrupted files, DICOMs with missing pixel array data, and unsupported manufacturing devices and monitors if the image metadata contains all the information required by the algorithm. For example, AI models for mammography interpretation may not be able to process images if view/projection (Cranio-Caudal (CC) or Medio-Lateral Oblique (MLO)) or laterality (left or right breast) information is missing from DICOMs. ClinValAI thoroughly analyzes the data to identify such inconsistencies and features a comprehensive input auditing mechanism to ensure a seamless external validation study.

2.3.2. *Data Standardization*

Standardizing inputs before initiating AI inferencing is necessary if there is variation from multiple data sources or if a data source requires enrichment before algorithmic processing can take place. ClinValAI's data standardization mechanism analyzes the findings of the input auditing logic and provides the feature of customizing the associated Linux Docker image to achieve data standardization and ensure the quality of the study data.

For our validation study of Mirai's performance, if a set of DICOMs is corrupted or missing pixel array information, the standardization mechanism does not pass them through the scoring workflow. Similarly, it removes images that do not match the study criteria – for example, deleting all the non-mammography images to ensure that only the acceptable modalities are included.

One of the important aspects of ClinValAI's data standardization mechanism is its ability to impute missing information. For example, if an image does not have laterality or projection information in the DICOM headers, the framework populates the DICOM metadata using the details from the manifest files. Moreover, if the required data is not available in the manifest file, it parses other descriptive DICOM headers to look for specific information for imputation. For example, AI algorithms for mammography interpretation expect laterality information in one of the *ImageLaterality*, *Laterality*, or *FrameLaterality* headers and projection information in the

ViewPosition header. If these tags are missing, ClinValAI's data standardization mechanism analyzes other subjective headers like *SeriesDescription* to systematically impute laterality and projection information into their respective tags. Thus, ClinValAI can be customized to facilitate effective data cleaning and preprocessing, information imputation, and data standardization.

2.4. *Scoring Workflow*

ClinValAI's scoring workflow (Figure 4) is the second stage in the framework. It comprises a data restructuring, an algorithm inferencing, and a score extraction mechanism.

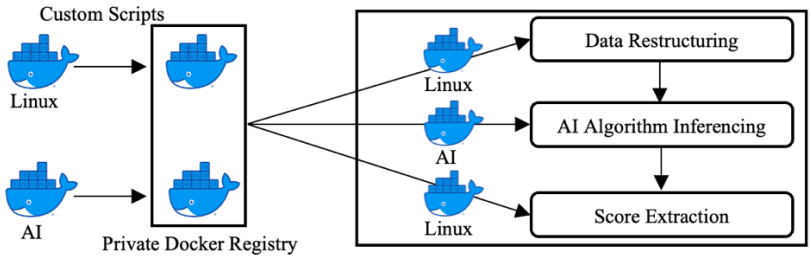

Fig. 4. ClinValAI's scoring workflow ensures consistent inferencing and systematic troubleshooting.

2.4.1. *Data Restructuring*

Various health institutions can organize patient data and medical images in different formats, requiring datasets to be systematized by patient ID, accession numbers, or date of collection. Different AI models can have their specific input structuring requirements. For example, a model may require all images from a patient to be in a single folder, while another may need additional sub-folders based on exam ID or modality. Different models may need varying numbers of images per exam – for instance, Mirai needs four standard 2D mammograms (CC and MLO views of the left and right breast), whereas some models can function even with unilateral exams. Some models can raise errors if inputs contain multiple images of the same view and laterality combination, while others can successfully score them. Moreover, some models can process 2D and 3D images simultaneously, while others can leverage separate Docker images depending on shape and modality. ClinValAI supports extensive data restructuring by enabling the customization of Docker images to account for model-specific variations by holistically analyzing the DICOM metadata and pixel array information, thereby establishing consistency between input and model criteria.

2.4.2. *AI Algorithm Inferencing*

ClinValAI enables effective customization of AI algorithms' Docker images to facilitate accurate scoring of exams. The Docker file is specified with the required environment variables and necessary scoring scripts, and the updated Docker image is used to spin up the AI model's Docker container to execute algorithmic processing. Information about the computational requirements of the AI algorithms can be utilized to identify the appropriate compute instances to be specified in our framework. To work with asynchronous inferencing workflows, our framework also features a polling mechanism depending on the inference time of each algorithm to ensure that the compute instances are not stalled due to inconsistent data, node failures, or other issues. Furthermore, our framework provides the flexibility of incorporating additional steps, such as drafting a list of input

studies to be processed or creating corresponding output folders for storing final results, depending on the models' prerequisites. Similarly, using our framework, we reconfigure Mirai's open-source Docker image via programmatic steps to streamline its inferencing workflow. Thus, by facilitating multiple customization features, ClinValAI enables robust validation of AI algorithms.

2.4.3. *Score Extraction*

After the completion of the scoring process, the model's generated files need to be processed to retrieve specific outputs of interest, such as image-, exam-, or patient-level scores. Different AI algorithms have different ways of representing outputs. ClinValAI enables customizing the Linux Docker image to follow the modes and steps to extract scores from diverse formats – from flat files like comma-separated values (CSV) documents to highly nested DICOM Structured Reports (SRs) and JavaScript Object Notation (JSON) objects. Similarly, ClinValAI also facilitates the storage of supplementary files, such as annotations in processed images or heat maps, and associated model explanations, if available, to facilitate improved interpretation for radiologists. Moreover, this step also records and organizes logs specific to the algorithm and workflow. Thus, ClinValAI facilitates systematic troubleshooting, effective scoring, and rigorous clinical validation of AI algorithms.

2.5. *Output Processing Workflow*

ClinValAI's output processing workflow (Figure 5) is the third and final stage in the framework. It comprises an output auditing and an output aggregation mechanism.

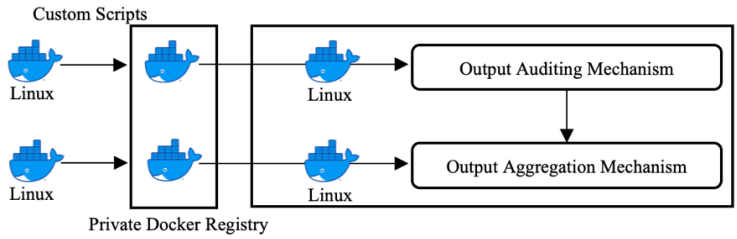

Fig. 5. ClinValAI's output processing workflow helps identify and analyze samples with missing results and aggregates final outputs for downstream analysis.

2.5.1. *Output Auditing Mechanism*

Once the scoring workflow has been executed, ClinValAI performs the essential task of verifying if results have been produced for all the inputs and if the generated files comply with the algorithm's expected outputs. Moreover, the framework facilitates examining if the required numeric values of interest, inference reports, and supplementary files can be extracted from the resulting outputs. ClinValAI identifies samples with missing output data, irretrievable scores, and corrupted output files to enable analysis of samples to be re-scored. If no outputs are generated for a patient's exam, infrastructure-specific logs can be inspected to check for issues related to compute instances or customization of the Docker images. If scores cannot be extracted from the model's output for an exam, algorithm-specific logs can be analyzed to check for inconsistencies and errors. Overall, ClinValAI facilitates a holistic output auditing mechanism for the streamlined validation of models.

2.5.2. *Output Aggregation Mechanism*

Outputs from individual workflows are hierarchically stored based on set number, batch number, and exam ID. Analyzing the complete dataset in the distributed format of cloud storage can be cumbersome. Before statistical analysis can be performed, ClinValAI systematically aggregates relevant details by appending all results to a relational database. After the completion of scoring workflows for all standardized batches of exams, the pipeline connects to the database and hierarchically uploads data from the audited results, supplementary files, and logs by inserting rows for every set, batch, and exam as demonstrated by the entity relationship diagram (Figure 6). During statistical analysis, this database is pulled to analyze findings. ExamIDs and Study UIDs (Unique Identifiers) are used to cross-reference the AI algorithm's results and the attributes of interest.

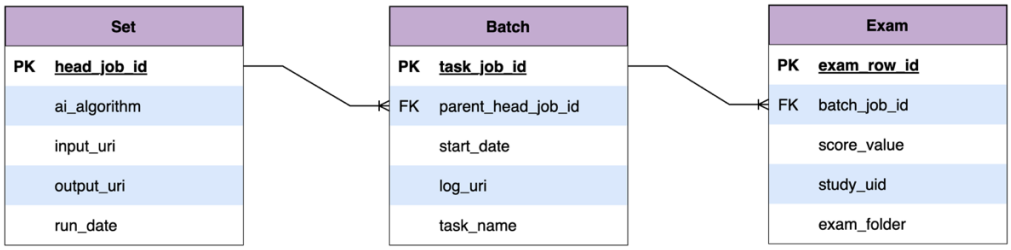

Fig. 6. Entity Relationship Diagram for the aggregated output data. Abbreviations: URI=Uniform Resource Identifier; PK=Primary Key; FK=Foreign Key

Thus, ClinValAI features multiple customizable workflows to establish optimized infrastructures for robust clinical validation of AI algorithms for medical imaging applications.

3. Evaluation and Results

To evaluate the utility of our framework, we used our ClinValAI-based cloud infrastructure to perform a rigorous external validation of Mirai, a state-of-the-art deep learning algorithm that predicts future breast cancer risk across five years by processing the four standard views of a 2D digital mammogram – Cranio-Caudal and Medio-Lateral Oblique views of the left and right breast.

3.1. *Patient Cohort*

All mammography screening exams from 2010-2014 performed across four imaging facilities in the University of Washington (UW) Medicine health system were reviewed for eligibility. Exams of women with age < 40 or ≥ 80 years, a personal history of breast cancer, or the presence of breast implants were excluded. Cancer outcomes at year 5 after every exam were collected via linkage to the Washington State cancer registry, which captures all breast cancers diagnosed within the state of Washington through December 31st, 2020, allowing for robust ground truth for all screening exams. Information on breast density and patient demographics, including age at the time of imaging and race, were obtained from the University of Washington Medicine electronic medical records. ClinValAI excluded exams with insufficient 2D screening images and processed 26,449 exams from 14,291 patients to generate Mirai scores. A total of 543 exams (2.1%) were followed by a breast cancer diagnosis within five years (88 in year 1, 92 in year 2, 112 in year 3, 119 in year 4, and 132 in year 5). Table 1 shows the patient characteristics. BI-RADS[25] categories 'heterogeneously dense'

and ‘extremely dense’ correspond to dense breasts, and ‘almost entirely fatty’ and ‘scattered fibroglandular’ correspond to non-dense breasts.

Table 1. Patient characteristics at each exam.

		Breast Cancer within 5 years	
	All	Yes	No
Variable	(n = 26,449)	(n = 543)	(n = 25,906)
Age			
40-49	7,014 (26.5%)	114 (21.0%)	6,900 (26.6%)
50-59	9,431 (35.7%)	151 (27.8%)	9,280 (35.8%)
60-69	7,082 (26.8%)	171 (31.5%)	6,911 (26.7%)
70-79	2,922 (11.0%)	107 (19.7%)	2,815 (10.9%)
Race			
White	20,365 (82.6%)	460 (87.1%)	19,905 (82.5%)
Black	1,649 (6.7%)	31 (5.9%)	1,618 (6.7%)
Asian	2,394 (9.7%)	33 (6.2%)	2,361 (9.8%)
Other	241 (1.0%)	4 (0.8%)	237 (1.0%)
Unknown	1,800	15	1,785
Breast density			
Not dense	11,659 (44.1%)	216 (39.8%)	11,443 (44.2%)
Dense	14,786 (55.9%)	327 (60.2%)	14,459 (55.8%)
Unknown	4	0	4

Values are number (%).

3.2. *Statistical Analysis*

A mammography exam was used as the unit of analysis. Nonindependence of multiple exams from the same women was accounted for in calculations of 95% confidence intervals (CIs) and p-values by using generalized estimating equations (GEE) or the nonparametric bootstrap, clustered by woman[26]. The Mirai algorithm provides cumulative risk predictions for years 1-5 following the index examination. The outcome used for evaluating the performance of Mirai was the presence/absence of a cancer diagnosis at each timeframe. The discrimination performance of Mirai was evaluated using receiver operating characteristic (ROC) curves, the area under the ROC curve (AUC), and Uno’s concordance index (c-index) as an overall summary over the 5-year timeframe[27]. The calibration of Mirai was evaluated using calibration plots and corresponding summaries of overall calibration (calibration-in-the-large) and the calibration slope[28]. To help distinguish between breast cancer detection vs. risk prediction performance, we performed the analyses using all available exams and then repeated the analyses after excluding exams that had a breast cancer diagnosis within six months. All statistical analyses were conducted using R (version 4.3, R Foundation for Statistical Computing, Vienna, Austria). All hypothesis tests were two-sided, with statistical significance defined as $p < 0.05$.

3.3. *Discrimination Performance*

AUCs ranged from 0.81 (95% CI: 0.75-0.86) for 1-year cancer outcomes with the 1-year Mirai scores to 0.70 (95% CI: 0.0.67-0.72) for 5-year cancer outcomes with the 5-year Mirai scores when including all examinations (Figure 7, Table 2). The c-index was 0.70 (95% CI: 0.67-0.72). After excluding 70 exams with a cancer diagnosis within six months, the AUC was 0.72 (95% CI: 0.56-

0.84) at 1 year and 0.68 (95% CI: 0.65-71) at 5 years, while the c-index was 0.68 (95% CI: 0.65-0.70). These values were more similar to previously reported results in other cohorts[19,29] after applying the same type of exclusion (Table 2)[29], though they were still on the lower end of the range.

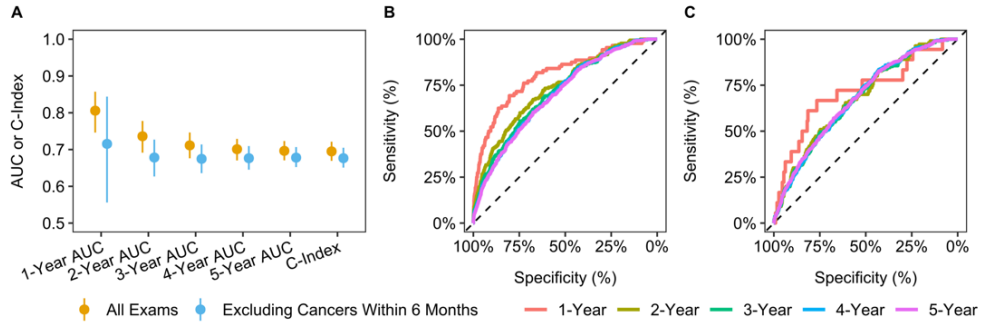

Fig. 7. Discrimination performance of Mirai. Panel A: ROCAUC values over time and the overall c-index. The orange values are based on all exams, and the blue values are after excluding cancers within six months. Error bars represent 95% CIs. Panel B: ROC curves at different time points based on all exams. Panel C: ROC curves at different time points after excluding cancers within six months.

Table 2. Discrimination performance of Mirai in the University of Washington and 7 previously reported cohorts.

	1-Year AUC (95% CI)	5-Year AUC (95% CI)	C-index (95% CI)
All Exams			
University of Washington, USA	**0.81 (0.75-0.86)**	**0.70 (0.67-0.72)**	**0.70 (0.67-0.72)**
MGH, USA[19]	0.84 (0.80-0.87)	0.76 (0.73-0.79)	0.75 (0.72-0.78)
Novant, USA[29]	0.78 (0.73-0.84)	0.75 (0.70-0.80)	0.75 (0.70-0.80)
Emory, USA[29]	0.83 (0.81-0.86)	0.76 (0.74-0.79)	0.77 (0.75-0.79)
Maccabi-Assuta, Israel[29]	0.86 (0.81-0.91)	0.75 (0.71-0.79)	0.77 (0.73-0.81)
Karolinska, Sweden[19]	0.90 (0.89-0.92)	0.78 (0.76-0.80)	0.81 (0.79-0.82)
CGMH, Taiwan[19]	0.90 (0.87-0.93)	0.79 (0.75-0.82)	0.79 (0.76-0.83)
Barretos, Brazil[29]	0.89 (0.86-0.93)	0.82 (0.78-0.86)	0.84 (0.81-0.88)
Excluding Cancers within 6 Months			
University of Washington, USA	**0.72 (0.56-0.84)**	**0.68 (0.65-0.71)**	**0.68 (0.65-0.70)**
MGH, USA[19]	0.71 (0.60-0.84)	0.71 (0.68-0.75)	0.69 (0.66-0.73)
Novant, USA[29]	N/A	0.72 (0.66-0.79)	0.72 (0.66-0.79)
Emory, USA[29]	0.74 (0.66-0.84)	0.71 (0.68-0.74)	0.69 (0.66-0.72)
Maccabi-Assuta, Israel[29]	N/A	0.68 (0.62-0.74)	0.70 (0.64-0.76)
Karolinska, Sweden[19]	N/A	0.71 (0.69-0.73)	0.71 (0.69-0.74)
CGMH, Taiwan[19]	0.84 (0.72-0.99)	0.70 (0.66-0.75)	0.70 (0.66-0.75)
Barretos, Brazil[29]	0.87 (0.80-0.94)	0.75 (0.70-0.80)	0.78 (0.74-0.83)

MGH = Massachusetts General Hospital; CGMH = Chang Gung Memorial Hospital.

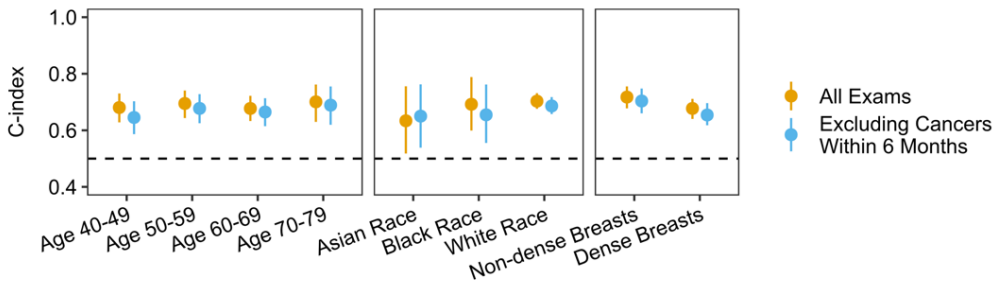

Fig. 8. Discrimination performance of Mirai within subgroups. Error bars: 95% CIs. Dashed line: AUC = 0.50.

Discrimination, as measured by the overall c-index, was also examined within subgroups defined by age, race, and breast density, as shown in Figure 8. There were no statistically significant differences in the c-index between subgroups (unadjusted $p > 0.094$ for each comparison).

3.4. *Calibration Performance*

Calibration plots for Mirai risk predictions versus observed at different timeframes are shown in Figure 9. The corresponding metrics of overall calibration (observed risk minus mean predicted risk) and the calibration slope are shown in Table 3. When all exams are included, the metrics indicated significantly overestimated risk in years 1-2 (overall calibration: -0.15% to -0.10%, $p < 0.014$ for both), but that Mirai was overall reasonably well calibrated for the later years, where the 95% CIs for overall calibration included zero (no difference between observed and predicted risk on average) and the 95% CIs for the calibration slope included 1 (predictions were not more or less extreme [farther from the mean] than observed on average).

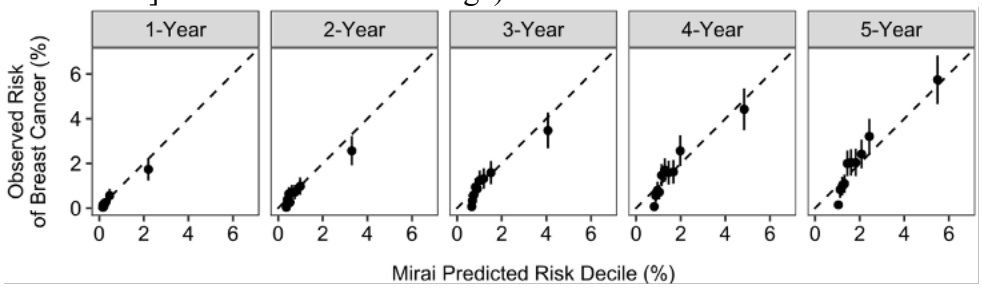

Fig. 9. Calibration plots of Mirai with all exams included, with predicted risks grouped into deciles of approximately equal size. Error bars represent 95% CIs. The dashed line corresponds to perfect calibration (intercept = 0, slope = 1).

Table 3. Calibration statistics for Mirai.

	All Exams				After Excluding Cancers within 6 Months			
	Overall Calibration*		Calibration Slope†		Overall Calibration*		Calibration Slope†	
Timeframe	Estimate (%)	(95% CI)	Estimate	(95% CI)	Estimate (%)	(95% CI)	Estimate	(95% CI)
1-year risk	-0.10	(-0.17, -0.03)	0.83	(0.52, 1.22)	-0.36	(-0.39, -0.32)	0.05	(0.00, 0.12)
2-year risk	-0.15	(-0.25, -0.04)	0.76	(0.51, 1.06)	-0.40	(-0.48, -0.32)	0.21	(0.08, 0.36)
3-year risk	-0.13	(-0.29, 0.03)	0.89	(0.60, 1.20)	-0.38	(-0.53, -0.26)	0.40	(0.20, 0.63)
4-year risk	-0.05	(-0.25, 0.14)	0.90	(0.63, 1.20)	-0.31	(-0.49, -0.12)	0.50	(0.27, 0.76)
5-year risk	0.09	(-0.15, 0.33)	1.03	(0.75, 1.32)	-0.16	(-0.38, 0.07)	0.66	(0.44, 0.91)

*Observed risk minus mean predicted risk; a value > 0 indicates the prediction under-estimated risk on average, and a value < 0 indicates the prediction over-estimated risk.

†A well-calibrated model has a calibration slope of 1; slope > 1 indicates that high predictions tended to underestimate risk (not high enough) and low predictions tended to overestimate risk (not low enough); slope < 1 indicates predictions tended to be more extreme than observed (high values too high and low values too low).

When exams with cancer diagnoses within six months were excluded, the calibration metrics substantially worsened (Table 3). Overall, Mirai significantly overestimated risk, more so at earlier timeframes (overall calibration: -0.40% to -0.36% in years 1-2 and -0.31% to -0.16% in years 4-5), and the calibration slopes were significantly less than 1 at all timeframes (calibration slopes: 0.05 to 0.66, $p < 0.012$ across years).

Thus, ClinValAI enabled the establishment of an effective cloud infrastructure to successfully perform the clinical validation of Mirai on a large and diverse dataset to study its generalizability.

4. Discussion

We introduce ClinValAI to promote the external clinical validation of AI algorithms on medical imaging exams, thereby providing the opportunity to reliably understand their real-world performance in healthcare settings and their impact on patient care, health, and safety. Our framework can be leveraged to evaluate the generalizability of deep learning models on healthcare data from diverse demographics to analyze the differences in performance across various sub-populations and identify biases. ClinValAI can facilitate the detection of models' failure modes and enable an understanding of AI's potential to function as a standalone tool for diagnostic applications.

An important consideration while using our work is the requirement to specify necessary programmatic steps while configuring Docker images to execute individual mechanisms. However, ClinValAI's multiple customizable features enhance its usability for validating AI models.

Our presented analysis is limited to just one deep learning algorithm. As a next step, we plan to leverage ClinValAI to perform a rigorous external validation study of Mirai and three commercial AI algorithms for breast cancer risk prediction on a large and diverse dataset of $\geq$ 40,000 mammograms from seven registries affiliated with the Breast Cancer Surveillance Consortium (BCSC). Utilizing our framework for this study will enable the evaluation of model performance at the woman, exam, and tumor levels, facilitating a comprehensive assessment of the generalizability of AI models. While we showcase ClinValAI's usability for medical imaging models, our work can be extended to validate AI models for various biomedical data modalities.

Finally, ClinValAI is equipped to provide opportunities to periodically retest performance. Vendors can analyze performance based on detailed findings from the results communicated by our framework. This encourages the development of explainable models to better reason performance, thereby enhancing the potential of receiving clinicians' trust. The streamlined feedback mechanism can support targeted algorithm fine-tuning efforts. This can foster enhanced academic-industry partnerships. The continuous monitoring feature enables analyzing variations in model performance vis-à-vis data drift and model drift. Overall, ClinValAI can pave the way for studying the capabilities of AI algorithms in optimizing clinical workflows and reducing the burden on the medical fraternity. ClinValAI's codebase and scripts for statistical analysis can be accessed here: https://github.com/OjasRamwala/ClinValAI.

5. Conclusion

The rise in commercial AI algorithms in clinical medicine and the associated generalizability concerns make rigorous validation indispensable to the clinical translation of AI tools. ClinValAI addresses critical challenges associated with external validation efforts and provides an easily customizable and cloud-agnostic framework to build scalable infrastructures to audit and monitor AI algorithms. By enabling large-scale external validation efforts on data from diverse cohorts, our work has the potential to foster health equity and overcome health disparities by promoting the development of robust, interpretable, and generalizable AI algorithms for healthcare applications.

Acknowledgments

This work was funded in part by the National Cancer Institute (grants P01CA154292, R01CA262023, R37CA240403, and R37CA292399), the American Cancer Society (grant 21-078-01-CPSH), the University of Washington Institute of Medical Data Science Pilot Award, an Amazon Web Services Health Equity Award, and the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

References

1. Parmar, S. K., Ramwala, O. A. & Paunwala, C. N. Performance Evaluation of SVM with Non-Linear Kernels for EEG-based Dyslexia Detection. in *2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)* 1–6 (2021). doi:10.1109/R10-HTC53172.2021.9641696.
2. Ramwala, O. A., Paunwala, C. N. & Paunwala, M. C. GRU-Based Parameter-Efficient Epileptic Seizure Detection. in *Biomedical Signal and Image Processing with Artificial Intelligence* (eds. Paunwala, C. et al.) 73–86 (Springer International Publishing, Cham, 2023). doi:10.1007/978-3-031-15816-2_4.
3. Fatemi, M. *et al.* Inferring spatial transcriptomics markers from whole slide images to characterize metastasis-related spatial heterogeneity of colorectal tumors: A pilot study. *J. Pathol. Inform.* **14**, 100308 (2023).
4. Schopf, C. M. *et al.* Artificial Intelligence-Driven Mammography-Based Future Breast Cancer Risk Prediction: A Systematic Review. *J. Am. Coll. Radiol.* (2023).
5. Ramwala, O. A., Dhakecha, S. A., Ganjoo, A., Visiya, D. & Sarvaiya, J. N. Leveraging Adversarial Training for Efficient Retinal Vessel Segmentation. in *2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* 1–6 (2021). doi:10.1109/ECAI52376.2021.9515093.
6. COVID-19 Diagnosis from Chest Radiography Images using Deep Residual Network | IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9225521.
7. Mulchandani, H. *et al.* Tonsillitis based Early Diagnosis of COVID-19 for Mass-Screening using One-Shot Learning Framework. in *2020 IEEE 17th India Council International Conference (INDICON)* 1–6 (2020). doi:10.1109/INDICON49873.2020.9342371.
8. Novel Multi-Modal Throat Inflammation and Chest Radiography based Early-Diagnosis and Mass-Screening of COVID-19. *Open Biomed. Eng. J.* **15**, 226–234 (2021).
9. Dalal, P. *et al.* Throat Inflammation Based Mass Screening of Covid-19 on Embedded Platform. in *Soft Computing and its Engineering Applications* (eds. Patel, K. K., Garg, D., Patel, A. & Lingras, P.) 277–288 (Springer, Singapore, 2021). doi:10.1007/978-981-16-0708-0_23.
10. Levy, J. *et al.* Artificial Intelligence, Bioinformatics, and Pathology: Emerging Trends Part I—an Introduction to Machine Learning Technologies. *Adv. Mol. Pathol.* **5**, e1–e24 (2022).
11. Levy, J. *et al.* Artificial Intelligence, Bioinformatics, and Pathology: Emerging Trends Part II––Current Applications in Anatomic and Molecular Pathology. *Adv. Mol. Pathol.* **5**, e25–e52 (2022).
12. Yang, X. *et al.* A large language model for electronic health records. *Npj Digit. Med.* **5**, 1–9 (2022).
13. Wu, S. *et al.* Deep learning in clinical natural language processing: a methodical review. *J. Am. Med. Inform. Assoc. JAMIA* **27**, 457–470 (2019).

14. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms | Breast Cancer | JAMA Network Open | JAMA Network. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2761795.
15. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
16. Maleki, F. *et al.* Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. *Radiol. Artif. Intell.* **5**, e220028 (2022).
17. Hsu, W. *et al.* External Validation of an Ensemble Model for Automated Mammography Interpretation by Artificial Intelligence. *JAMA Netw. Open* **5**, e2242343 (2022).
18. Ramwala, O. A. *et al.* Establishing a Validation Infrastructure for Imaging-Based Artificial Intelligence Algorithms Before Clinical Implementation. *J. Am. Coll. Radiol.* (2024) doi:10.1016/j.jacr.2024.04.027.
19. Yala, A. *et al.* Toward robust mammography-based models for breast cancer risk. *Sci. Transl. Med.* **13**, eaba4373 (2021).
20. Mooney, S. J. & Pejaver, V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annu. Rev. Public Health* **39**, 95–112 (2018).
21. Alternative models for sharing confidential biomedical data | Nature Biotechnology. https://www.nature.com/articles/nbt.4128.
22. Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction | Journal of the American Medical Informatics Association | Oxford Academic. https://academic.oup.com/jamia/article/27/9/1393/5868591.
23. Voss, K., Gentry, J. & Auwera, G. V. D. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. (2017) doi:10.7490/F1000RESEARCH.1114631.1.
24. miniwdl — miniwdl documentation. https://miniwdl.readthedocs.io/en/latest/.
25. D'Orsi, C. J., Sickles, E. A., Mendelson, E. B. & Morris, E. A. *2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. (American College of Radiology, 2014).
26. Huang, F. L. Using Cluster Bootstrapping to Analyze Nested Data With a Few Clusters. *Educ. Psychol. Meas.* **78**, 297–318 (2018).
27. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**, 1105–1117 (2011).
28. Crowson, C. S., Atkinson, E. J. & Therneau, T. M. Assessing Calibration of Prognostic Risk Scores. *Stat. Methods Med. Res.* **25**, 1692–1706 (2016).
29. Yala, A. *et al.* Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **40**, 1732–1740 (2022).