

A Technical Performance Study and Proposed Systematic and Comprehensive Evaluation of an ML-based CDS Solution for Pediatric Asthma

Shauna M. Overgaard, Ph.D.¹, Kevin J. Peterson, Ph.D., M.S.¹ Chung H. Wi, M.D.^{2,3}, Bhavani Singh Agnikula Kshatriya^{1,4}, Joshua W. Ohde, Ph.D.¹, Tracey Brereton, M.S.¹, Lu Zheng, Ph.D.¹, Lauren Rost, Ph.D.¹, Janet Zink, Ph.D.¹, Amin Nikakhtar, Ph.D., M.S.¹, Tara Pereira¹, Sunghwan Sohn, Ph.D.⁴, Lynnea Myers, Ph.D., R.N., Ph.N.³, Young J. Juhn, M.D.^{2,3}

¹Center for Digital Health, Mayo Clinic, Rochester, Minnesota;

²Precision Population Science Lab, Mayo Clinic, Rochester, Minnesota;

³Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, Minnesota;

⁴Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester

Abstract

Achieving optimal care for pediatric asthma patients depends on giving clinicians efficient access to pertinent patient information. Unfortunately, adherence to guidelines or best practices has shown to be challenging, as relevant information is often scattered throughout the patient record in both structured data and unstructured clinical notes. Furthermore, in the absence of supporting tools, the onus of consolidating this information generally falls upon the clinician. In this study, we propose a machine learning-based clinical decision support (CDS) system focused on pediatric asthma care to alleviate some of this burden. This framework aims to incorporate a machine learning model capable of predicting asthma exacerbation risk into the clinical workflow, emphasizing contextual data, supporting information, and model transparency and explainability. We show that this asthma exacerbation model is capable of predicting exacerbation with an 0.8 AUC-ROC. This model, paired with a comprehensive informatics-based process centered on clinical usability, emphasizes our focus on meeting the needs of the clinical practice with machine learning technology.

Introduction

Clinical care of chronic diseases requires a personalized treatment approach, and the heterogeneity of variables contributes to the challenges of managing chronic diseases such as asthma, the most common chronic condition in children and adolescents in the United States (1-3). A major impediment to optimal asthma care is the lack of efficient and effective clinical decision support (CDS) systems that enable high-value asthma care and quality outcomes while reducing clinician burden (4-6) and costs (7-9). While risk stratification is challenging at the point of care because it relies on the synthesis of large heterogeneous information, there is strong potential for the use of machine learning, natural language processing, and deep learning algorithms to identify and stratify patients by symptom severity and socioeconomic status (10, 11). In pediatric asthma management, efficient and effective review of electronic health records (EHR) and timely clinical decisions are impeded by the varied location, contents and formatting of features characterizing pediatric asthma, given that asthma may serve as an endpoint to a variety of underlying pathologies (12). Furthermore, relevant patient data for asthma management is often stored as unstructured free-text narratives or coded as non-uniform (standardized) inputs into multiple EHR workflows (13). If providers are to better manage uncontrolled asthma and to optimize asthma care, efficient and effective use of EHRs is critical.

Our aim is to develop an AI solution to support clinicians through 1) predicting future risk of asthma exacerbation for their patients (risk stratification and resource management), 2) providing this risk evaluation in the context of a summary of relevant information for asthma management (reduction of EHR review burden), and 3) offering actionable intervention options. We also aim to evaluate the developed algorithm in the context of a comprehensive AI implementation and evaluation framework developed by Park et al. 2020 (14), and further bolster this framework by applying our internally developed model documentation framework.

Methods

Through the support and collaboration of the Mayo Clinic Center for Digital Health, Precision Population Science

Lab and the Department of Pediatric and Adolescent Medicine Artificial Intelligence (DPAM-AI) program, a multidisciplinary team of practicing clinicians, data scientists, and translational informaticians was formed to develop, evaluate, and execute a machine learning-based CDS solution to offer augmented artificial intelligence that optimizes asthma care.

User Needs and Initial Workflow Assessment

The tool will be used by physicians, nurse practitioners, nurses, and care coordinators/asthma managers from the Division of Community Pediatric and Adolescent Medicine and Department of Family Medicine for outpatient asthma management in children. Providers aim to efficiently identify pediatric patients with poorly controlled asthma or with moderate to severe asthma and improve their asthma care by applying precision asthma clinical care in both the personal and social context. To represent the primary user group of the CDS tool, clinical stakeholders are central to all developmental processes and engaged on a weekly basis to participate in shaping and prioritizing translational informatics deliverables that are people-centric, provide value, and are evidence based. Currently, the clinical workflow involves the clinician reviewing the EHR prior to their upcoming appointments to better understand the clinical background and current health status for each patient. For patients with asthma, this requires searching multiple locations within the EHR to determine the patient's asthma status and what potential interventions, if any, need to be considered. The process of locating all asthma related data is challenging and time consuming, reducing time with patients and their families. The translational informatics team reviewed feasibility study data previously conducted by the DPAM-AI group (15-17). Follow-up interviews were then conducted to create an initial list of user requirements for workflow integration, visual presentation, and explainability of the model output. The user requirements specification was evaluated against the NASSS (Non-Adoption, Abandonment, Scale-up, Spread, and Sustainability) framework (18) to assess requested features for sustainability, and scalability, a method employed to increase the likelihood of CDS tool value and broad adoption. The consensus was to create an AI-assisted CDS tool imbedded within the current workflow to help clinicians review the EHR efficiently and effectively, prioritizing resources for high-risk patients. Importantly, the clinical stakeholders required that the custom AI/ML algorithm identify patients at high risk for asthma exacerbation and further curate a list of monitorable asthma-related measures for comprehensive and efficient tracking and review by providers.

System Description

The purpose of the comprehensive CDS tool is to offer augmented artificial intelligence by reducing the amount of time a provider spends locating any single patient's features in the EHR that may contribute to asthma evaluation and intervention. By requirements of clinical stakeholders, the ML/AI algorithm is one feature in the CDS tool and will present a risk score to indicate the likelihood that a patient will experience an asthma exacerbation within the next twelve months. This *comprehensive* information would provide insight for the autonomous clinician to use in preparing a diagnosis and care intervention. Therefore, the Mayo Clinic's Asthma-Guidance and Prediction System (A-GPS) is being developed as an AI-assisted CDS tool that extracts patient-level data to a) predict the likelihood that the patient will experience an asthma exacerbation (AE), b) explain the predicted risk by detailing the relevant factors contributing to the predicted likelihood of AE, c) contextualize the algorithm by placing the outcome within a visualization of relevant clinical information pertaining to a patient's asthma status, and d) operationalize the algorithm by offering actionable, timely interventions (precision asthma care).

We anticipate that the improved workflow will involve reviewing the A-GPS tool by clinicians immediately before each patient appointment. Prior to entering the room, clinicians often review the EHR of the patient in the upcoming appointment. This process currently requires significant time to locate and review the patient's background, recent visits, medications, and current diagnosis. To improve this process, A-GPS will be integrated into the current workflow allowing clinicians to access asthma-related information in one location, enabling an efficient review of the patient's clinical information prior to an appointment. The ability to use the A-GPS tool without changing the clinician's current workflow is a high priority among users and stakeholders.

Technical Performance Evaluation

Technical performance evaluation will be assessed and validated through the data card system developed by Mayo Clinic. We also conducted an evaluation of the predictive model component of A-GPS. Development of this machine learning model focused on the prediction of AE risk for a patient, where AE risk is defined as the probability of having at least one exacerbation within the next 12 months. Patient is labelled as having exacerbation based on three specific criteria: a) An inpatient/hospitalization visit for asthma diagnosis, b) An ED visit for asthma diagnosis, c) An outpatient visit of a patient with asthma diagnosis along with usage of oral corticosteroids medications. If a patient meets any one of the three criteria, then the patient is labelled as having asthma

exacerbation. A rule-based NLP algorithm was developed and validated to identify exacerbation patients with any of the above criteria as described in Model Features section.

Cohort Selection

Mayo Clinic Rochester Campus patients between the ages of 6 and 17 with active asthma were selected for this study, where active asthma is defined as at least one previous Mayo Clinic clinical visit with an associated asthma diagnosis in the past twelve months.

Model Features

Currently, there are 28 variables which include both structured and unstructured data that are used to train the models. All of these features are identified as clinically significant factors by the domain experts which assist in determining the probability of having an asthma exacerbation. Both unstructured and structured variables have Yes, No, and/or Unknown labels. The structured data are retrieved from the Mayo database using customized SQL queries, but the unstructured features are extracted using rule-based natural language processing algorithms developed under MedTagger IEarchitecture (19) (developed specifically to address corresponding asthma condition). For example, NLP was used to identify pre-determined the asthma criteria of the patient, which is one of the 28 features in the A-GPS study. Each NLP algorithm was developed using hand-crafted rules with clinical concepts suggested by domain experts and was validated against the Manual Chart reviewed dataset as a gold standard.

Data Quality Evaluation

Many machine learning metadata systems are focused on experimental aspects (hyperparameters, F1 score for a particular run, etc.) (20, 21). While these are important considerations, healthcare data can be complex/challenging, which in turn determines performance of AI systems and provides insights into explainability, transparency, and accountability of AI systems. The Mayo Clinic Data Card is a component of the organization's comprehensive model documentation tool, with a goal to incorporate a system that can be shared, re-used, and clinically validated. The data card includes relevant project model information, is organized by feature, and contains both structured and unstructured data. It also consists of a definition and category for each feature, data provenance such as sample query and data source, distribution of data availability, distribution of comorbidities aligned with clinical criteria, clinical description, and interpretation. A data card for each of the 28 model features was processed for the interdisciplinary team's evaluation and feedback.

Model Evaluation

Model evaluation was conducted using 5-fold cross-validation. Evaluation metrics selected for the binary dependent variable included the area under the receiver operating characteristic curve (AUC-ROC) and F1 score. For F1 evaluation of the binary output, a 0.5 positive/negative threshold was initially chosen. The final threshold will be determined via analysis of the clinical application and through feedback from care providers.

Results

Model Selection

The following candidate models were developed from several different architectures: logistic regression, support vector machine, random forest, Gaussian Naive Bayes, and multilayer perceptron. Each model was implemented using the scikit-learn Python machine learning framework (22).

Overall, logistic regression model outperformed the other candidates producing a 0.8 AUC-ROC. F1-score is shown in Table 1 (Figure 1).

Table 1. Logistic Regression model output

Threshold	Mean F1
0.1	0.421527
0.2	0.522789
0.3	0.531575
0.4	0.484955
0.5	0.456748
0.6	0.430991
0.7	0.4161
0.8	0.388956
0.9	0.339465

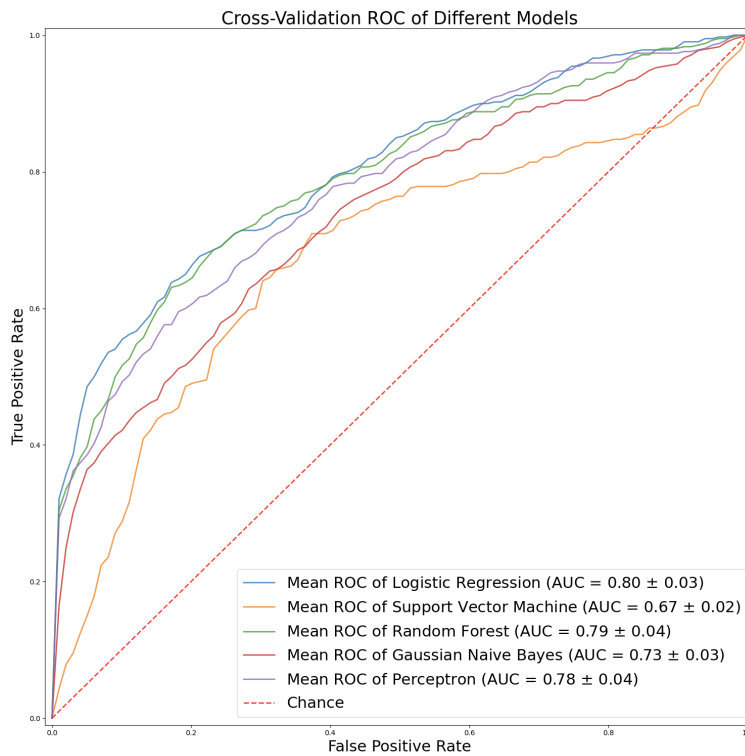


Figure 1. Comparison of AUC score of various models

Final parameters chosen for the logistic regression model were: ($C=0.23357214690901212$, $max_iter=10000$, $penalty='l1'$, $solver='liblinear'$). All other parameters of logistic regression were applied with default values defined in Scikit-learn package (22).

Discussion

We have described our proposed machine learning-based CDS system to alleviate the clinician burden of consolidating relevant information for optimal care of pediatric asthma patients and presented the methods and results from our study of algorithm performance to predict future risk of asthma exacerbation. To deploy our tool safely and effectively in the clinical setting, we further propose a systematic and comprehensive evaluation plan.

AI Evaluation Plan

Our plan is to conduct an AI evaluation using the framework described by Park et al. 2020, which includes five

phases (Figure 2) to address studies of technical performance, usability and workflow, and health impact (14). We iteratively follow our model documentation steps throughout the AI evaluation to ensure dissemination based on our experience and a literature study, which will accompany the tool throughout its life cycle and into maintenance.

Model Documentation

We followed our standardized model documentation framework developed internally, which includes five steps: (1) Prepare, (2) Develop, (3) Validate, (4) Deploy, (5) Maintain. Components from the TRIPOD, ML Test Score, CONSORT-AI, Model Card reporting guidelines, and internal best practices prompt model documentation items (23-26). Our model documentation framework serves as a reference to inform future phases and stakeholders, and accompanies the AI/ML tool for knowledge continuity through evolution and maintenance. By promoting transparency and standardizing documentation requirements, we position our ML-based CDS tool for translation, address and mitigate bias and risk proactively, and communicate use cases and deployment strategies across all stakeholder groups.

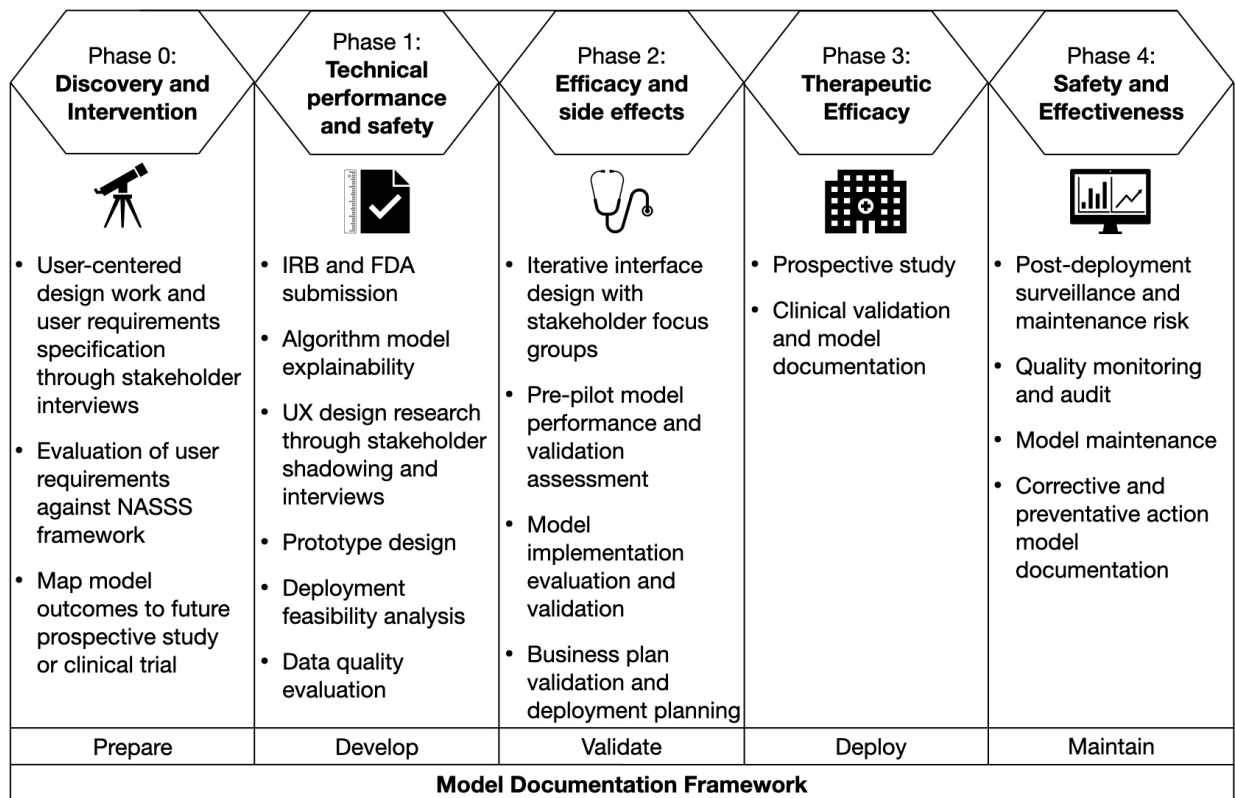


Figure 2. Phased research framework for evaluation of AI applied to the A-GPS project based on Park et al. (2020), with the internally developed model documentation framework (14).

Phase 0: Discovery and Intervention

Clinical stakeholders are central to the CDS tool and AI/ML developmental processes and are currently engaged on a weekly basis to participate in shaping and prioritizing translational informatics and data science deliverables. While we have completed Phase 0, user needs and workflow assessment will continue to inform model development and translation. In addition to traditional aspects of model development, collaborative translational informatics and data science processes include design thinking, process improvement, and implementation science to ensure the AI/ML-augmented CDS tool is people-centric, provides value, and is evidence based (27).

User Needs and Initial Workflow Assessment

As a tool to support asthma management in outpatient practice, the CDS tool users are composed of a variety of clinician roles, such as pediatricians, primary care providers, pulmonologists, asthma care coordinators and nurses. Preparation and evaluation of user requirements specifications through workflow assessment and clinician interviews indicated a consensus to create an AI-assisted CDS tool (i.e., A-GPS) to help clinicians review the EHR efficiently and effectively, prioritizing resources for high-risk patients. Importantly, the clinical stakeholders required that the custom AI/ML algorithm identify patients at elevated risk for asthma exacerbation and further curate a list of monitorable asthma-related measures for comprehensive and efficient tracking and review by providers in a single dashboard view. Moreover, the current clinical workflow will not change with the integration of A-GPS. Clinicians require access to the A-GPS tool at the same time they would typically review the EHR.

Identification of Target Outcomes and Key Performance Indicators

To prepare for clinical evaluation, model outcomes were included in a broader prospective evaluation to measure clinician satisfaction and the feasibility of the AI-CDS tool implementation. The plan will assess the capacity to execute the workflow triggered by the model and estimate the potential impact of healthcare delivery factors and work capacity constraints. In a future version of the tool, we plan to integrate a remote patient monitoring spirometry device and embed asthma patients' home-based data in the CDS tool for proactive asthma management. As primary users of the spirometry device, the satisfaction and requirements of pediatric patients and their caregivers will be fully evaluated through in-depth interviews.

Model Documentation: Prepare

We explained the medical context (including whether diagnostic or prognostic) and rationale for developing the AE prediction model, specifying the objectives and key elements of the model through the use of TRIPOD (23). TRIPOD was also used to establish eligibility criteria for participants, data sources, extraction dates, and description and distribution of cohorts. The intended use of the AI intervention in the context of the clinical pathway and intended users was explained through CONSORT-AI (25). We identified out-of-scope use cases using model cards (26).

Phase 1: Technical Performance and Safety

Continued algorithm modeling performance optimization will be scheduled before medical use setting evaluation. UX design specialized translational informaticians will engage with practice components to iteratively construct prototype and evaluate usability and workflow to determine effectiveness, efficiency, satisfaction, ease of use, explainability, and utilization.

IRB and FDA Submission

Organization-specific Institutional Review Board (IRB) processes will be followed in accordance with standard protocols. Mayo Clinic's legal department will be engaged to determine whether the algorithm falls under the purview of Food and Drug Administration (FDA) oversight and fits within regulatory definitions.

AI/ML Model Explainability

Literature surrounding the current state of AI/ML-based CDS documentation recognizes inconsistencies in standards and governance surrounding explainability. The ineffective balance of CDS tool intelligence with explainability, creates gaps in translation, implementation, and accountability (28). The lack of sufficient documentation further proves challenging for stakeholder adoption calling for the development of enhanced strategies to effectively translate AI-CDS tool integration into clinical settings (29).

The translational informatics and data science teams will engage clinical stakeholders in creating materials to ensure appropriate interpretation, contextual information, and educational support for use of the model. When it is known that the ML-based CDS tool functions technically, its fit into the clinical workflow must be evaluated and education and documentation must be provided to explain the algorithm and its limitations to effectively translate between the perspectives of experts that create and support the technology and the perspectives of experts that employ the solution to patients. Evaluating the interpretation needs of the clinicians, preferences for display of model output (e.g., percentage vs. binary threshold), and feature contributions will be assessed. Concurrently, the teams will engage clinician stakeholders in the development of model documentation to support explainability(30). Strategic efforts to promote explainability include the application of a ML-based CDS tool documentation framework grounded in scientific research addressing known challenges by encompassing interdisciplinary best practice reporting

requirements that follow phases of model development (Prepare, Develop, Validate, Deploy, Maintain) for knowledge continuity throughout the solution's life cycle.

UX Research and Prototype Design

Continued UX design research includes clinician shadowing and interviews to understand user requirements and facilitate optimal workflow integration, estimate potential impact of healthcare delivery factors, and work capacity constraints on achieved benefit. UX design research will be conducted through collection of qualitative and quantitative data to identify all user groups, understand user needs, probe for optimal tool design to support clinical decision making and routine workflow for each group. The data will be collected from a minimum of 12 clinicians through interviews. Based on research goals, four formats of UX data collection methods will be used: (1) in-person shadow, (2) virtual stakeholder interview, (3) virtual user interview, (4) screen capture of EHR walkthrough. These four methods provide comprehensive contextual data from different facets. The shadowing will allow for in-situ investigation of facility workflow, patient flow and physician-patient interaction for decision-making during visits. The interviews and EHR walkthrough will focus on problem-probing and user needs consolidation.

Feasibility Analysis

The interdisciplinary team will plan and document how the model will be deployed into the environment. This will include review of architecture requirements, a diagram of architecture supporting the model, technical specifications of model features, clear determination of variable extraction, mapping, and methods of incorporation into model. In addition to the technical review of the model, the team will review potential bottlenecks and solutions, identify success metrics, establish expected output of discovery and usability tests, evaluate feasibility of the algorithm's implementation, and identification of elements required for capture once algorithm has been deployed in end-user practice domain.

Data Quality Evaluation: AI Metadata Management

While the Mayo Clinic Data Card framework is an initial step towards transparent and comprehensive, consolidated model documentation management, the longer-term vision includes a fully vetted model-agnostic tool for managing all Mayo AI-related artifacts. As a bottom-up system model, it will house reusable components, or cards, at the feature, model, and project level. These components all have a one-to-many relationship with one another and are designed to be accessed via a user-friendly interface with version control, and to be searchable, discoverable, executable, and auditable. Data governance and provenance are also defining characteristics of the tool, and it is rooted in the FAIR initiative principles of digital asset findability, accessibility, interoperability, and reusability (31).

Model Documentation: Develop

We defined the outcome that is predicted by the model and described how predictors were used in developing and validating the model (TRIPOD). We specified model type, model-building procedures, methods for internal validation, and all measures used to assess model performance. Performance measures for the model were reported as well. Relevant and evaluation factors, model performance measures, decision thresholds, variation approaches, preprocessing, and training data were all identified. We provided a lay summary of model development based on Mayo Clinic Translational Informatics Best Practices. Intersectional results investigating algorithm discriminatory behavior against subgroups and intersectional subgroups were detailed. Participant characteristics, including demographics, clinical features, and available predictors were described. Detailed testing was also conducted for inclusion and reproducibility considerations. In addition, we described the results of risk assessment tests, detailed a risk management plan, and outlined user requirements based on Mayo Clinic Translational Informatics Best Practices.

Phase 2: Efficacy and Side Effects

The CDS tool interface will be designed iteratively with clinical stakeholders. The algorithm will undergo a controlled performance and efficacy evaluation by intended users in the medical setting. This will first incorporate model evaluation on EHR data trained in silent mode. Following validation and approvals, a pilot study will be administered on a subset of pediatric providers to assess usability and workflow integration of the CDS tool.

Interface Design

UX design specialized translational informaticians will employ agile methodology to iteratively design the AI/ML-augmented CDS tool with the clinician focus group. The team will build the interface and test workflows to support

the seamless integration of the algorithm in end-user practice. Relevant systems and back-end workflow will be mapped to identify key systems impacted by CDS tool integration. The user workflow and risks will be designed and validated to compare current and future state, given CDS tool impact. A change management strategy will be prepared to assist in clinicians' transition in order to minimize workflow impact and maximize value of the CDS tool.

Pre-Pilot Model and Business Plan Validation

Following initial statistical assessment of model performance, the model will be configured to train on EHR data extracts in silent mode. Model performance after training will be validated. Following model completion and the verification of infrastructure needs, the team will construct the business integration plan, review the implementation timeline, training, and communication materials, and seek approvals from the department and enterprise governance committees. Results of model performance, validation, and proposed implementation timeline will be reviewed by the interdisciplinary team, practice proponents, and EHR analysts for consensus approval to initiate pilot in a clinical setting. Adhering to internal protocols, the team will submit a change request for implementation into the EHR.

Pre-Post Implementation Evaluation

Study items for clinician satisfaction and time saved with current asthma management tools/interface will be drafted based on findings from analysis of interview data and information extraction time. To evaluate the impact of the new A-GPS patient summary tool implementation systematically, a baseline evaluation will be captured prior to any changes made to interface or workflow, and repeated following implementation of the A-GPS summary tool. Outcomes will be measured and organized using Proctor's Implementation Outcomes. Implementation outcome metrics include use frequency and prevalence of the tool. Service outcome metrics can be clicks, screen transitions, and time of information retrieval process. Client outcome metric would be satisfaction level of users. Success in meeting user requirements will be evaluated by tool users themselves and recorded following implementation.

Pilot Model and Business Plan Validation

A pilot study evaluating the effectiveness of the algorithm, its ability to correctly process input data to generate accurate, reliable, and precise output data, and the association of the algorithm's output and targeted clinical indicators will be conducted. Meanwhile, validation utilities will provide insight into model performance for the organization's patient population. Important model performance metrics like true-positive, true-negative, false-positive, false-negative rate, and AUC-ROC will be calculated against testing data. The ML-based CDS tool will undergo an internal evaluation for potential commercialization.

Model Documentation: Validate

We will present the full prediction model to clinicians in order to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline values), demonstrate how to use the prediction model, specify the output of the AI intervention, specify whether there was human-AI interaction in the handling of the input data and what level of expertise was required of users, and explain how the AI intervention's output contributed to decision-making or other elements of clinical practice (23, 25).

Phase 3: Therapeutic Accuracy

A clinical validation study will evaluate the algorithm in its intended purpose to produce clinically meaningful, precise, accurate, and reliable output for the target clinical indicators. A randomized controlled trial will be executed and will identify adverse events, effectiveness, efficiency, satisfaction, ease of use, learnability, and utilization.

Prospective Study/Randomized Clinical Trial

A randomized controlled trial will demonstrate the clinical validity, efficacy, and safety of the algorithm. A parallel-group, non-blinded, dual-site, 2-arm pragmatic randomized controlled trial will be planned for a period of 9 months (2-month enrollment period followed by a 6-month trial) to assess the feasibility of implementation of the CDS tool within primary care practices 30 dyads of primary care clinician-patient with asthma will be recruited from Mayo Clinic Rochester and Mayo Clinic Health System primary care sites. Clinicians under intervention group after randomization will access the CDS tool that can support asthma management whereas those under control group will follow usual care for asthma. Primary end point is the change of clinicians' and caregivers' satisfaction score with asthma care during the trial. Also, we will assess if the CDS tool may improve efficiency and quality of asthma care by measuring time to review patient's electronic health records as well as direct interaction time with patients at the point of care. The CDS tool can automatically extract and synthesize large amounts of pertinent patient data related to asthma management from electronic health records to support clinicians and care teams to optimize asthma care

through the following: 1) provision of a one-page summary of the most relevant clinical information for asthma management for the past 3 years, 2) prediction for future asthma exacerbation risk within a year from a machine learning algorithm, and 3) offering actionable, timely and guided interventions (precision asthma care). After the trial, we will review model performance, evaluate usability data, consider, and document adjustments and needs.

Model Documentation: Deploy

Even in the era of electronic health records, there is lack of effective and efficient CDS tool for streamlining asthma management (7-9). While many CDS tools using machine learning algorithms have been developed and even deployed within electronic health records, few CDS tools have been tested via randomized controlled trial (32). We plan on conducting a clinical trial in which the results may include, but not limited to, safety, clinical outcomes, care quality, care cost, and satisfaction of end users (e.g., clinicians, patients, and caregivers). Once clinical validation is demonstrated and the tool is deployed to health care institutions, the model performance overall as well as in subgroups needs to be monitored on a regular basis to prevent potential bias.

We aim to give an overall interpretation of model results, considering objectives, limitations, and results from similar studies and other relevant evidence, as well as discuss the potential clinical use of the model and implications for future research (23). Methods of integration for the AI intervention into clinical settings, including any onsite or offsite requirements and training, will be described (25). Model limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses, will be detailed (25). We will detail a prospective study for technical validation of the model, aiming to communicate whether the model correctly processes input data to generate accurate, reliable, and precise output data and validates model performance for clinical application (Mayo Clinic Translational Informatics Best Practices).

Phase 4: Safety and Effectiveness

Post-deployment surveillance and post-production management will include an evaluation of adoption and sustainability of the AI algorithm utilization.

Post Deployment & Maintenance Risk

The performance of the algorithm and relevant processes will be tracked and compliance with relevant regulatory standards and requirements ensured. The algorithm performance will be monitored for quality assurance testing, patient safety, and incorrect results. Conditions will be re-evaluated to determine whether the algorithm is exposed to external factors that may affect its accuracy including but not limited to evolving societal norms and values, covariate shift, downtime risk, and degradation of model. The need for algorithm retraining, algorithm improvement, and algorithm retirement will be assessed at a minimum of every six months.

Quality Monitoring and Audit

Tracking mechanisms for optimal and high-quality performance will be established, including frequent status updates to leadership, meta-data monitoring, algorithm failures, and distributional shift. The internal audit team will be engaged for quality monitoring and regulatory compliance.

Maintenance

The interdisciplinary team will troubleshoot algorithm related issues or failures, the addition of new features, and versioning, updating the model documentation as needed. The internal operations team will review technical issues related to algorithm performance or user feedback.

Model Documentation: Maintain

We aim to discuss any limitations of the model (such as nonrepresentative sample, few events per predictor, missing data) and provide information about the availability of supplementary resources associated with the model. We will ensure and document considerations taken in regard to data, human life, bias mitigation, risks and harms, and appropriate use cases. Contact information communicating where to send questions, comments, or additional

concerns about the model will be provided. We aim to describe a quality assurance plan, address the potential of external and internal drift, develop incident monitoring logs, and provide contact information for accountable teams.

Conclusion

Our aim was to describe an ML-based CDS tool for pediatric asthma management developed by our multidisciplinary team of clinicians, data scientists, translational informaticians, and UX experts at Mayo Clinic, present the results of our algorithm study, and propose our comprehensive evaluation framework. This tool has the potential to reduce the clinician's EHR review burden by gathering features to optimize precise pediatric asthma management and to predict the likelihood of asthma exacerbation with an 0.8 AUC-ROC. The goal is to provide this risk evaluation in the context of a summary of relevant information for asthma management and offer actionable intervention options. The construction and application of the phased research approach requires strategic partnership across our multidisciplinary team and stakeholders, and a commitment to rigorous evaluation. As we progress this work, we will continue to leverage a published comprehensive evaluation framework on which we will infer our problem-specific studies. This evaluation plan will be further enhanced through application of our internally developed model documentation framework to ensure rigorous evaluation, transparency, and knowledge continuity at each step: 1) Prepare, 2) Develop, 3) Validate, 4) Deploy, 5) Maintain. These efforts are intended to support the potential for future adoption of this tool on a national and international level.

References

1. Prevention CfDCA. Asthma 2021 [Available from: <https://www.cdc.gov/healthyschools/asthma/index.htm>.
2. Valet RS, Gebretsadik T, Carroll KN, Wu P, Dupont WD, Mitchel EF, et al. High asthma prevalence and increased morbidity among rural children in a Medicaid cohort. *Ann Allergy Asthma Immunol*. 2011;106(6):467-73.
3. Druss BG, Marcus SC, Olfson M, Tanielian T, Elinson L, Pincus HA. Comparing the national economic burden of five chronic conditions. *Health Aff (Millwood)*. 2001;20(6):233-41.
4. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. *J Am Med Inform Assoc*. 2019;26(2):106-14.
5. Holmgren AJ, Downing NL, Bates DW, Shanafelt TD, Milstein A, Sharp CD, et al. Assessment of Electronic Health Record Use Between US and Non-US Health Systems. *Jama Intern Med*. 2021;181(2):251-9.
6. Linzer M, Bitton A, Tu SP, Plews-Ogan M, Horowitz KR, Schwartz MD, et al. The End of the 15-20 Minute Primary Care Visit. *J Gen Intern Med*. 2015;30(11):1584-6.
7. Shiffman RN, Freudigman M, Brandt CA, Liaw Y, Navedo DD. A guideline implementation system using handheld computers for office management of asthma: effects on adherence and patient outcomes. *Pediatrics*. 2000;105(4 Pt 1):767-73.
8. McCowan C, Neville RG, Ricketts IW, Warner FC, Hoskins G, Thomas GE. Lessons from a randomized controlled trial designed to evaluate computer decision support software to improve the management of asthma. *Med Inform Internet Med*. 2001;26(3):191-201.
9. Bell LM, Grundmeier R, Localio R, Zorc J, Fiks AG, Zhang X, et al. Electronic health record-based decision support to improve asthma care: a cluster-randomized trial. *Pediatrics*. 2010;125(4):e770-7.
10. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol*. 2020;145(2):463-9.
11. Juhn YJ, Beebe TJ, Finnie DM, Sloan J, Wheeler PH, Yawn B, et al. Development and initial testing of a new socioeconomic status measure based on housing data. *J Urban Health*. 2011;88(5):933-44.
12. von Mutius E, Smits HH. Primary prevention of asthma: from risk and protective factors to targeted strategies for prevention. *Lancet*. 2020;396(10254):854-66.
13. Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. *Yearb Med Inform*. 2014;9:14-20.
14. Park Y, Jackson GP, Foreman MA, Gruen D, Hu J, Das AK. Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open*. 2020;3(3):326-31.
15. Seol HY, Rolfes MC, Chung W, Sohn S, Ryu E, Park MA, et al. Expert artificial intelligence-based natural language processing characterises childhood asthma. *BMJ Open Respir Res*. 2020;7(1).
16. Seol HY, Shrestha P, Muth JF, Wi CI, Sohn S, Ryu E, et al. Artificial intelligence-assisted clinical decision support for childhood asthma management: A randomized clinical trial. *PLoS One*. 2021;16(8):e0255261.

17. Seol HY, Sohn S, Liu H, Wi CI, Ryu E, Park MA, et al. Early Identification of Childhood Asthma: The Role of Informatics in an Era of Electronic Health Records. *Front Pediatr.* 2019;7:113.
18. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *J Med Internet Res.* 2017;19(11).
19. Liu H, Bielinski SJ, Sohn S, Murphy S, Waghlikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. *AMIA Jt Summits Transl Sci Proc.* 2013;2013:149-53.
20. Lee Y, Ragguett RM, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *J Affect Disord.* 2018;241:519-32.
21. Arefeen MA, Nimi ST, Rahman MS, Arshad SH, Holloway JW, Rezwan FI. Prediction of Lung Function in Adolescence Using Epigenetic Aging: A Machine Learning Approach. *Methods Protoc.* 2020;3(4).
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011;12:2825-30.
23. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *Br J Surg.* 2015;102(3):148-58.
24. Breck E, Cai S, Nielsen E, Salib M, Sculley D, editors. The ML test score: A rubric for ML production readiness and technical debt reduction. 2017 IEEE International Conference on Big Data (Big Data); 2017: IEEE.
25. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, Spirit AI, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ.* 2020;370:m3164.
26. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*; Atlanta, GA, USA: Association for Computing Machinery; 2019. p. 220–9.
27. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. *NPJ Digit Med.* 2020;3:107.
28. Addressing Ethical Dilemmas in AI: Listening to Engineers Report [updated 2021. Available from: <https://standards.ieee.org/initiatives/artificial-intelligence-systems/ethical-dilemmas-ai-report.html>
29. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise Qc. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak.* 2020;20(1):310.
30. Cuttillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, et al. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med.* 2020;3:47.
31. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018.
32. Angus DC. Randomized Clinical Trials of Artificial Intelligence. *JAMA.* 2020;323(11):1043-5.