ELSEVIER

Integrating Artificial Intelligence and Machine Learning Into Cancer Clinical Trials

John Kang,* Amit K. Chowdhry,† Stephanie L. Pugh,‡ and John H. Park§,‖

The practice of oncology requires analyzing and synthesizing abundant data. From the patient's workup to determine eligibility to the therapies received to the post-treatment surveillance, practitioners must constantly juggle, evaluate, and weigh decision-making based on their best understanding of information at hand. These complex, multifactorial decisions have a tremendous opportunity to benefit from data-driven machine learning (ML) methods to drive opportunities in artificial intelligence (AI). Within the past 5 years, we have seen AI move from simply a promising opportunity to being used in prospective trials. Here, we review recent efforts of AI in clinical trials that have moved the needle towards improved prediction of actionable outcomes, such as predicting acute care visits, short term mortality, and pathologic extranodal extension. We then pause and reflect on how these AI models ask a different question than traditional statistics models that readers may be more familiar with; how then should readers conceptualize and interpret AI models that they are not as familiar with. We end with what we believe are promising future opportunities for AI in oncology, with an eye towards allowing the data to inform us through unsupervised learning and generative models, rather than asking AI to perform specific functions.

Semin Radiat Oncol 33:386–394

Introduction

In just a few years, we have seen artificial intelligence (AI) and machine learning (ML) in healthcare transition from buzzwords to clinical application. AI/ML has permeated not just the frontlines of mammography,[1,2] stroke,[3] sepsis,[4] readmission,[5] acute kidney injury,[6,7] diabetic retinopathy[8,9] and melanoma detection[10] but more controversially and less visibly, the murky backlines of health economics like prediction of no-shows[11] and healthcare utilization.[12]

The main theme of existing AI in medicine is diagnosis. Can AI also be used to improve outcomes? In this paper, we explore several such examples where prediction of acute events in prospective trials are used to improve longer term outcomes for patients. We highlight how AI is complementary to traditional statistics. We also discuss the philosophy of AI in clinical studies and how to incorporate this information into our understanding of traditional statistics paradigms that are the foundation of clinical research. We end with a discussion of the future of AI in trials, and explore how AI can move beyond prognostication and decision making into generation of knowledge.

*Department of Radiation Oncology, University of Washington, Seattle, WA.
†Department of Radiation Oncology, University of Rochester, Rochester, NY.
‡American College of Radiology, NRG Oncology Statistics and Data Management Center, Philadelphia PA.
§Department of Radiation Oncology, Kansas City VA Medical Center, Kansas City, MO.
‖Department of Radiology, University of Missouri Kansas City School of Medicine, Kansas City, MO.
Conflicts of interest: None.
Address reprint requests to John Kang, Department of Radiation Oncology, University of Washington, 1959 NE Pacific St, Box 356043, Seattle, WA 98195. E-mail: johnkan1@alumni.cmu.edu

Present Oncology AI Trials

Early AI-driven studies in oncology have focused on retrospective model training and validation to predict cancer control and toxicity, similar to a biomarker.[13,14] In recent years, AI has been to either drive decision points in phase 2 or phase 3 clinical trials (Table 1).

Emergency Room Visits and Hospital Admissions

Hong et al. at Duke University sought to tackle the problem of acute care visits, which include unplanned emergency department (ED) and hospital admissions during radiation. Admissions are problematic oncologically as it may interrupt radiation treatments, allowing accelerated repopulation to

https://doi.org/10.1016/j.semradonc.2023.06.004

Table 1 Examples of AI-driven Prospective Clinical Trials

Trial Name/Dates	Trial Design	Model Methods	Model Objective	Clinical Action or Arms	Clinical Objective
SHIELD-RT[15] (2019)	Randomized quality improvement	Gradient boosted trees	Predict if patient has >10% risk of acute care visit	Twice-a-week clinic visit vs. standard once-a-week	Decrease acute care visits during RT
University of Pennsylvania, Manz et al. [16,17] (2019)	Prospective silent → phase 3 cluster wedge	Gradient boosted trees	Predict 6-month mortality	Weekly email on high-risk patients w day-of text message vs. usual care	Increase serious illness conversation rate
INRT-AIR[18] (2019-2022)	Phase 2	Hybrid radiomics and 3D CNN	Predict individual neck node positivity risk	Involved nodal radiation to suspicious nodes	Risk of solitary elective nodal recurrence
Stanford, Gensheimer et al. [19] (2020-2021)	Quality improvement	Cox proportional hazards with structured and unstructured variables	Predict survival over time	Weekly emails with patients predicted to survive <2y	Increase advance care planning rate
DARTBOARD[20] (2022-)	Phase 2 randomized	Same as INRT-AIR ± AI-driven ART deformable registration and planning	CBCT image segmentation via proprietary CNN architecture	INRT +/- daily ART (near marginless)	Decrease xerostomia

Abbreviations: ART, adaptive radiotherapy; CBCT, cone-beam CT; CNN, convolutional neural network; INRT, involved nodal radiotherapy; RT, radiotherapy.

occur and as well as increase healthcare costs. In the first paper, they first used a ML framework to train various models suited for biomedical data, ultimately selecting gradient boosted trees (GTB) whose predictive performance reached an area under the receiver operating characteristic curve (AUROC) of 0.798.[21]

Using this model, Hong and colleagues subsequently designed and ran a single institution randomized quality improvement trial to see if their GTB model could decrease the number of acute care visits. Out of 962 patients who received radiation between January-June 2019, the GTB model selected 311 patients at high risk (>10% predicted risk of acute care visit) who were randomized to either standard once-a-week clinic visits or mandatory twice-a-week visits during radiation.[15] For patients deemed low risk or high risk by the GTB model, the true rate of acute care visits was 2.7% and 17.4%, respectively. The randomized trial showed that in the high risk cohort, twice-a-week visits during radiation decreased the acute care visit rate from 22% to 12%, and in the window of 2 weeks after radiation from 33% to 22%. In this example, the GTB model classified patients as high- and low-risk and the between arm comparisons were analyzed using traditional statistical methods, showing how both analysis methods can be used in tandem.

Short Term Mortality

Predicting short term mortality is a popular goal, as this knowledge can be used to drive interventions such as prompting goals of care discussion or referrals to hospice and/or palliative care services. Oncologists are quite poor at predicting survival of their patients[22,23] and there are often delayed referrals to palliative care.[24]

Popular prognostic models include recursive partitioning analysis (RPA) and diagnosis-specific graded prognostic assessment (dsGPA). These decision-tree models use a limited number of variables such as age and performance status to estimate overall survival in different patient populations[25,26] with recent incorporation of molecular biomarkers.[27,28] Yet, the EHR holds a vast amount of data that is underutilized for prognosis estimation.

Parikh and colleagues at the University of Pennsylvania aimed to leverage EHR data to predict 180 day mortality to drive timely serious illness conversations (SIC) in patients with metastatic cancer. They performed model development[29] where they trained and internally validated random forest, gradient boosted trees (GBT), and logistic regression algorithms on a cohort of 26,525 patients and 559 features (following feature engineering and selection). At a pre-specified alert rate of 2% (ie, the proportion of patient encounters flagged), all 3 models were correct approximately 50% of the time when they predicted that a patient would die within 180 days (positive predictive value PPV ranging from 45% to 51%).

Manz et al. validated this GTB framework in a prospective silent trial that classified patients with new oncology encounters as either high or low risk of 6 month mortality.[16] They surveyed oncologists to fix the event rate at 2.5% (ie, what

proportion of total patients are predicted to be sick) and achieved a PPV of 45% (Table 3), which is similar to 49% achieved by the GTB model during the previous model development study.[29]

This silent trial was subsequently followed up by a randomized stepped-wedge cluster intervention trial.[17] Over 14,000 patients were enrolled from 8 clinic groups that were randomized to 4 intervention wedge periods. The groups received either a behavioral nudge—weekly email discussing SIC performance and up to 6 patients predicted to be at high risk of 6 month mortality (≥10%) with opt-out text messaging reminders—or usual care which was weekly email summarizing SIC performance. Across all encounters, the intervention increased SIC from 1.3% to 4.6%. In the subset of high risk encounters (~4100 patients), intervention increased SIC from 3.6% to 15.2%. It was not reported what the increase in SIC was in patients who died within 6 months (iem, truly had short term mortality or the ground truth for high risk).

Other AI models have incorporated both EHR structured data—diagnoses, procedures, vital signs, labs—as well as unstructured data such as clinical notes. Gensheimer and colleagues at Stanford used EHR, inpatient billing, and cancer registry data to train a Cox proportional hazards model to predict overall survival. They used over 12,500 patients with metastatic solid tumors and 4000 features including notes, labs, vital signs, procedures, and diagnoses.[30] Notes were represented as bags of words, where the top 100,000 1- to 2-word phrases were tallied for each note. This model was subsequently compared with physician prediction and a traditional performance status-based model, and shown to be superior.[31] In a follow up quality improvement trial to validate their model, Gensheimer et al. compared the rate of advance care planning (ACP) in a cohort of oncology clinics that received weekly emails of patients predicted to have <2 year survival to a control cohort of clinics without such emails. The intervention group had an ACP documentation rate of 35% compared to 3% in the control.[19]

Head and Neck: Extranodal Extension

Head and neck cancer represents one of the more challenging sites to deliver high doses of radiation due to nearby organs-at-risk which can cause very significant side effects including xerostomia and mucositis. Due to the better prognosis of HPV-associated oropharyngeal cancer, several completed and ongoing trials are looking to de-escalate therapy in this sub-population of patients to spare toxicity, with promising results.[32-35] While HPV is a clearly detectable biomarker, another prognostic marker is the presence of extranodal extension (ENE) in regional lymph nodes, which is an indication for postoperative chemoradiation.[36] If patients who will have ENE can be identified at diagnosis, they may potentially be offered definitive chemoradiation upfront and avoid the toxicity of additional surgery. However, identifying ENE can be very challenging for radiologists, with only about 50% of pathologic ENE cases detectable via imaging by head and neck neuroradiologists.[37] One of the major advancements in computer vision over the past decade was the rise of deep convolutional neural networks (CNN)[38] in an era of high performance computing.

Leveraging CNN towards an important clinical question, Kann and colleagues trained and internally validated a CNN model to detect ENE using 653 segmented lymph nodes from 270 patients.[39] As ground truth, a node was considered to have ENE if the pathology report (1) confirmed the presence of lymph node positivity and ENE (microscopic or macroscopic) and (2) it could be determined from the report the node's location, anatomic level, and size. Their CNN model achieved sensitivity, specificity, PPV and NPV of 0.88, 0.85, 0.66, and 0.95, respectively, to predict ENE.

Taking the next logical step, Kann et al. performed external validation on a separate institution's data as well as publicly available data from the The Cancer Genome Atlas head and neck imaging data, showing that their model exceeds the the diagnostic ability of 2 head and neck neuroradiologists.[40] This article led to editorials pointing out operational issues, chief among them (1) the extra effort required to segment lymph nodes to run the model (notable as other specialties do not perform this in practice); (2) that radiographically negative necks can hide occult positive nodes and/or ENE (which would not have been included in training the model); and (3) that model is learning to predict microscopic ENE mostly from non-radiologic features (making comparisons with expert radiologists perhaps unfair).[41-43]

To further demonstrate the generalizability of their CNN model to clinical trial data, Kann et al. applied it to data from the completed Phase III trial ECOG-ACRIN E3311,[33] which aimed to de-escalate therapy in patients with HPV-associated oropharyngeal cancer without high-risk pathologic criteria such as macroscopic (>1mm) ENE. Yet >30% of enrolled patients demonstrated ENE, requiring postoperative chemoradiation.[44] Using 311 segmented lymph nodes from 177 presurgery CT scans and pathology reports from ECOG-ACRIN E3311, they compared their model's performance against 4 head and neck radiologists who were provided with an educational tool to help diagnose ENE. On this high-quality, contemporary dataset, the algorithm achieved AUROC 0.85 and outperformed human experts.

Head and Neck: Involved Nodal Radiation

Another potential area for treatment de-escalation is in decreasing the dose or volume of the elective nodal fields. There is controversy about coverage of elective nodal regions during radiotherapy for head and neck cancer. Using a data-driven approach, Chen and colleagues at UT-Southwestern used data from the INFIELD trial of dose and volume de-escalation[45] to train a hybrid multi-objective radiomics and 3D-convolutional neural network model to predict for lymph node malignancy.[46] The model classifies each lymph node as involved or suspicious based on CT and PET imaging. This model was tested prospectively in the INRT-AIR (Involved Nodal RadioTherapy using AI-based Radiomics) trial.[47]

The goal of INRT-AIR was to eliminate elective neck treatment and focus only on involved or suspicious nodes in newly diagnosed oropharynx and larynx/hypopharynx squamous cell carcinoma. The gross disease and CTV were treated to 70 and 63 Gy, respectively, in 35 fractions with suspicious nodes treated to 66.5 Gy. The primary endpoint was the risk of solitary elective nodal recurrence (ie, nodal recurrence in the classic ENI field without synchronous in-field or distant failure). In 68 patients enrolled, initial results at 1 year showed no solitary elective nodal recurrences and 1 patient who died from in-field local progression.[18] Along with work by Kann et al. to detect ENE (discussed above), these early results in head and neck cancer show promise for AI-driven treatment de-escalation.

Tissue Biomarkers

One promising avenue for bringing AI into clinical trials is exploratory analysis of previously conducted trials to look for biomarkers.

Esteva and colleagues performed a secondary analysis of 5 phase III trials in prostate cancer comparing external beam RT with or without hormone therapy to determine if incorporating deep learning of digitized histopathology slides was superior to standard clinical risk stratification models.[48] Using a large pre-trained network,[49] the authors trained their deep learning models on approximately 5600 patients and 16,000 histopathology slides. Compared to NCCN risk group models using standard clinical variables, the authors showed that deep learning of histopathology images resulted in 0.92% to 14.6% relative improvement in predicting 6 binary outcomes: 5/10 year distant metastasis, 5/10 year biochemical failure, 10 year cancer specific survival, and 10 year overall survival.

Adaptive Radiotherapy

Adaptive radiotherapy, where one modifies the treatment plan due to anatomic changes or treatment response during treatment, holds significant promise to improve workflows and outcomes, with many practical considerations to be addressed.[50] Platforms such as Varian Ethos (Varian Medical Systems, Palo Alto, CA) incorporate proprietary AI-driven image segmentation, deformable registration, and planning.[51] Several Varian-funded trials are exploring Ethos in a variety of different disease sites with the potential to benefit from online replanning.[52] Given the promising results from INRT-AIR (discussed above), Sher and colleagues are using involved nodal radiation with or without near marginless daily adaptation on Ethos in DARTBOARD (Daily Adaptive Radiotherapy to Better Organ-at-Risk Doses).[20]

Key Points About Interpreting AI Trials

Diagnostic Testing and Model Evaluation

Diagnostic testing is traditionally used to determine the presence or absence of a disease or condition, or response from a therapy. In classification, a diagnostic test can be considered either positive or negative to predict if a patient has the disease, for example. Whether the patient has the disease (sick) or not (healthy) is the true state. There are various statistics that assess how well a binary diagnostic test performs to determine the patient's true state. Some of the more common ones are listed in Table 2. Sensitivity is the true positive rate, or the probability of a positive result given the patient is sick. Specificity, or true negative rate, is the probability of a negative result given the patient is healthy. Negative predictive value (NPV) is the probability of being healthy given a negative test while the positive predictive value (PPV) is the probability of being sick given a positive test, assuming a consistent prevalence of disease.

In classification, a tradeoff must be made between sensitivity and specificity. For a given diagnostic test, if the sensitivity increases, the specificity decreases, and vice versa. For example, if we would like to increase our ability to diagnose prostate cancer, we could decrease the prostate specific antigen (PSA) threshold for a biopsy from (say) 3.0 ng/mL to 0.5 ng/mL. This would absolutely increase the test's ability to detect prostate cancer in sick patients (sensitivity), though would also decrease the test's ability to rule out prostate cancer in healthy patients(specificity).

Many of the aforementioned examples utilizing AI in clinical trials are using these models similar to diagnostic tests to aid in selecting the best treatment option for patients. The Manz et al. model[27] predicted whether patients were low risk (survival >180 days) or high risk (survival ≤180 days). The NPV was 96.9% and PPV was 45.2%, meaning the probability of identifying a patient as healthy given the model classified the patient as healthy was 96.9% while the probability of identifying a patient as sick given the model classified the patient as sick was only 45.2%.

As seen in Table 3, the resultant specificity was very high (98.6%) while the sensitivity was low (27.4%). This discrepancy illustrates the tradeoffs that must be made to meet performance requirements: to be correct 45% of the time when predicting a high risk patient when the model is fixed to

Table 2 Commonly Used Binary Classification Metrics With Descriptions and Definitions

Classification Metric	Description	Definition
Sensitivity / recall / true positive rate	Proportion of sick patients correctly labeled	$\frac{TP}{TP+FN}$
Specificity / true negative rate	Proportion of healthy patients correctly labeled	$\frac{TN}{TN+FP}$
Precision / positive predictive value	Proportion of sick predictions correctly labeled	$\frac{TP}{TP+FP}$
Negative predictive value	Proportion of healthy predictions correctly labeled	$\frac{TN}{TN+FN}$

Abbreviations: FP, false positives; FN, false negatives; TP, true positives; TN, true negatives.

Table 3 Confusion Matrix Summarizing the Performance of Prospective Validation of the Penn Gradient Boosted Tree Model to Predict 180 Day Mortality[16]

Total Patients (24,582)	Predict High Risk (23,963)	Predict Low Risk (619*)	
High risk patient (1022)	742 false negatives	280 true positives	Sensitivity 27.4%
Low risk patient (23,560)	23,221 true negatives	339 false positives	Specificity 98.6%
	NPV 96.9%	PPV 45.2%	

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.
High risk and low risk patients' survival were >180d and ≤180d, respectively. The paper reported total patients, sick patients, healthy patients, negative predictive value, positive predictive value, and sensitivity. The rest of the values were derived with possible rounding errors.
* The number of "Predict sick" events was fixed at 619/24,582 = 2.5% via a survey of oncologists.

only allowing 2.5% of patients to be labeled as high risk, the model will miss approximately 75% of high risk cases.

Is there a way to help determine what thresholds to use or a way to compare diagnostic tests (ie, ML models) against each other? One way is through the receiver operating characteristic curve (ROC), which plots the sensitivity (true positive rate) by 1-specificity (false positive rate).[53] The area under the curve of a ROC (AUROC) allows one to compare how well a classifier or diagnostic test performs against other such models or tests. AUROC is a number from 0 to 1 in which 1.0 represents perfect prediction and 0.5 represents no discrimination (ie, the diagnostic test performs as well as a coin flip; Fig. 1). The AUROC can also be viewed as the average sensitivity across all possible false positive rates.

In model development for the Hong clinical trial, the authors compared random forest, support vector machine, logistic regression with LASSO regularization and the GTB model to classify patients as either high- or low-risk for acute care visits.[21] The GTB model had the highest AUROC of 0.798 and was chosen for their subsequent randomized trial[15] (Fig. 2).

AI Compared to Frequentist/Bayesian Analysis

In order to see how AI methods can improve clinical trials, it would be helpful to compare and contrast it with traditional frequentist and Bayesian methods. The overwhelming majority of clinical trials follow the frequentist paradigm of statistics that utilizes null hypothesis significance testing with the use of *P*-values, confidence intervals, along with type I and II error controls. "Statistical significance tests are part of a rich piecemeal set of tools intended to assess and control the probabilities of misleading interpretations of data—often called error probabilities".[55] In the context of a prospective, double blinded, randomized trial with pre-registration of endpoints, multiplicity adjustments, and checks to ensure model assumptions are valid, one can make a reliable statistical inference. Some of the weaknesses of frequentist trials stem from the need for an adequate sample size. When

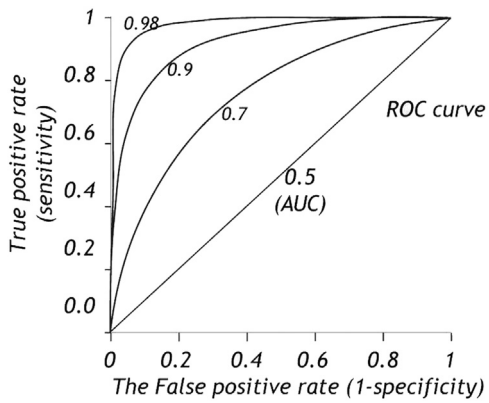

Figure 1 ROC-curves with different areas under the curve (AUC) or c-statistic. The better the discrimination, the larger the AUC or c-statistic. An AUC of 0.5 means no discrimination, an AUC = 1 means perfect discrimination. Figure from Christensen 2009.[54] (Color version of figure is available online.).

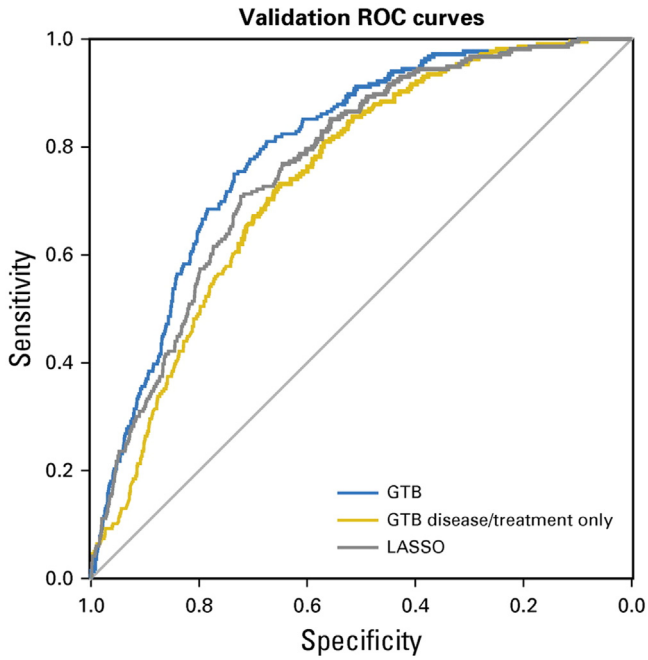

Figure 2 Validation receiver operating characteristic (ROC) curves for machine learning techniques. Although all 3 methods yielded strong predictive results, gradient boosted trees (GTB; 0.798) had greater area under the ROC curve than random forest (0.770; not shown), support vector machine (0.759; not shown), and least absolute shrinkage and selection operator (LASSO) logistic regression (0.768) methods. All had greater area under the ROC curve compared with GTB trained on only disease and treatment-related characteristics (0.742). Figure from Hong et al. 2018.[21] (Color version of figure is available online.)

enough patients are not recruited for a study, one cannot make a strong conclusion from the accrued patients. In addition, subgroup analyses are not definitive because they are typically not adequately powered, so a lot of information is not able to be utilized fully.[56]

Bayesian statistics is now becoming a more popular method with 2 randomized studies recently published,[57,58] but has been used in radiation oncology for over 30 years, as it was used in the Medical Research Council (MRC) neutron and CHART trials.[59] Bayesians approach data very differently; they use a prior distribution that uses available background information mathematically described in various distributions (eg, Gaussian, Beta, or Gamma). Instead of looking at alternatives to a null hypothesis, Bayesians calculate the conditional probability of the tested hypothesis given the prior information in combination with the most recent available data, such as those obtained in a clinical trial. Once the data is evaluated a statistical inference is made via the combination of what you believed before the trial (prior distribution), available data from the trial (likelihood), to form new beliefs (posterior distribution).

One of the strengths of Bayesian statistics—especially in its classical form—is that no information is wasted, as any data available will update your prior beliefs. Problems with Bayes include the difficulty of having an accurate prior (a long standing controversy) and the question of its ability to properly control for error probabilities in contrast to frequentist statistics.[60]

Machine learning is a field borne out of the intersection of computer science and statistics. The term "statistical machine learning" is nearly synonymous with "machine learning" as almost all ML models use statistics in some way (with notable exceptions such as many clustering methods). While there is a significant overlap in the statistical methods, there is an overall difference in approach. In general, a statistical inference approach makes assumptions on the distribution of the data and the model that best captures a phenomenon, whereas an ML approach makes minimal assumptions about the data and phenomenon (treating them as a black box) and aims to achieve a goal, such as predicting an outcome. These 2 approaches are well-summarized in Leo Breimen's "Statistical Modeling: The 2 Cultures" paper as the data modeling culture and the algorithmic modeling culture, respectively[61]; the lessons from this 2001 paper are still relevant today. ML is infamous for using layers of high-dimensional complex mathematical algorithms where the human user cannot understand how the predictions are being made, admittedly with recent progress on interpretable AI.[62]

As illustrated in the prior case examples, AI models are generally validated through several steps of validation on independent test sets, and then inferences are made based on the accuracy of the algorithm's predictions. AI then can leverage the strength of both frequentist and Bayesian analyses. ML can be subjected to randomized studies and be subject to the same rigorous tests as other interventions in a frequentist manner. AI methods can also preserve the information from the studies and use that data to further refine its algorithms in a Bayesian manner. They will be able to make predictions on subgroup analyses that can be further validated, such as in real world data studies, where "big data" evaluation can analyze thousands of data points with continuous refinement, a task that current clinical trials methods are unable to perform.

AI/ML should not be thought of as a replacement to traditional statistics. The multi-faceted uses of AI along with the different types of questions it answers allows for synergy with traditional statistics in clinical trials. Frequentist and Bayesian statistics are still appropriate for making inferences about data. AI algorithms and ML models are useful to classify patients and can be used to aid in determining eligibility or used in stratification of a clinical trial. Targeting interventions to certain groups of patients based on classification from an AI algorithm or ML model allow for improved and personalized treatment. Clinical trials may still be designed and analyzed using traditional frequentist or Bayesian approaches while implementing components of AI/ML, as shown in the above use cases.

The Future of AI in Oncology Trials

Clinical Trial Screening

Clinical trial recruitment is a challenging task. Departments often hire multiple staff members (clinical research coordinator) who will enroll, administer, and track patients. Automated systems using NLP may improve eligibility matching, decrease time spent, and increase enrollment.[63] NRG Oncology, an oncology research group funded by the National Cancer Institute, is moving towards utilizing NLP to aid participating institutions in screening. NRG-CC005 is a randomized cancer screening trial to determine if 10 year colonoscopies are non-inferior to 5 year colonoscopies in terms of cancer incidence. Because these patients may have had their first diagnosis of 1 to 2 nonadvanced adenomas within the prior 4 years, this study offers NLP to be used by participating institutions to search their EHR to identify potentially eligible patients.

Clinical Trial Design

Generative models which can synthesize data have come to the forefront of mainstream AI through ChatGPT (Chat Generative Pretrained Transformer). Blurring the line between discriminative and generative models, Liu and colleagues at Stanford sought to use data-driven methods to determine the impact of overly restrictive eligibility criteria in clinical trials in order to design more inclusive trials.[64] Using eligibility criteria from 10 trials for advanced non-small cell lung cancer (NSCLC), they designed rules to select patients from a real-world de-identified database from Flatiron Health (New York, NY) that would have met eligibility and also received the same treatments, using propensity scores to adjust for confounding and emulate randomization. Using synthetic clinical trials using various combinations of inclusion/exclusion criteria, they found that many of these criteria,

such as specific lab values, could be relaxed to increase the number of eligible patients with minimal effect on outcomes.

Clinical Trial Conduct

NRG Oncology is implementing AI models to conduct radiation plan treatment reviews with the goal to replace the manual review by a radiation oncologist or, if unable to complete the review, identify cases requiring a manual review. This would significantly reduce workload on the trial's radiation oncologists without compromising quality assurance. Currently this is an endpoint on NRG-GU009, a phase III randomized trial using prostate ribonucleic acid (RNA) expression to individualize concurrent therapy with radiation.

Uncertainty and Missingness

Although the promise of ML in medicine is great, as with any technology, there are pitfalls to be aware of. Well publicized ML issues include the shutting down of Google Flu Trends due to its faulty predictions[65] and allegations of racism when the COMPAS algorithm was used for determining recidivism rates during criminal sentencing.[66] David Watson, lecturer in AI at King's College points out a critical issue that "[h]igh-performance algorithms are often opaque, in the sense that it is difficult or impossible for humans to understand the internal logic behind individual predictions. This raises fundamental issues of trust. How can we be sure a model is right when we have no idea why it predicts the values it does?".[67]

For clinical trials evaluating and using machine learning, messy data, such as incomplete assessments, non-adherence of treatment, and missingness, can be problematic, as for traditional clinical trials. Techniques to handle missing data, such as imputation techniques are helpful, but do not provide a fix for all problems of missing data. Going forward, increasing attempts to prevent missing data, as well as sensitivity analyses to understand the impact of missing data on the results and subsequently, conclusions, are warranted.[68] Similarly, nonadherence to a treatment schedule is a problem for data analysis for virtually all randomized trials, which tend to analyze based on intention-to-treat, and appropriate causal inference methods are important to handle biases that may result from this problem.[69]

Emergent Logistics in Real World Deployment

The applications of clinical trial evidence are not always straightforward and AI-driven trials have a unique challenge of uncovering emergent issues related to new technology. Screening for diabetic retinopathy using deep convolutional neural network classification of images was one of the early examples of deep learning in health that rivaled the performance of physicians.[8] The model designed by Gulshan and colleagues at Google Research was trained on high-quality curated datasets. However, when Google Health prospectively validated this model in Thailand through a human-centered evaluation in nurse-run screening clinics, they ran into several unexpected real world issues. The real-world clinic images would contain artifacts not present in the training data and thus present triaging problems. Another issue emerged when the model recommended an ophthalmologist evaluation, and patients living in rural areas did not have the means or finances to travel long distances to specialists concentrated in urban area.[70] Much as how current clinical trials need to consider generalizability of treatments and patient eligibility, clinical trials using AI will need to consider emergent issues centering on the human evaluation of AI models.

Conclusion

The past, present, and future of AI in oncology trials is bright, but there is much work to be done. We have discussed the groundwork that has—and is—being laid by various research groups. Many challenges remain such as intersystem compatibility, data entry errors, and (not randomly) missing data, which need to be overcome, and we look forward to seeing these challenges faced in the years ahead.

References

1. Lo JY, Baker JA, Kornguth PJ, et al: Computer-aided diagnosis of breast cancer: Artificial neural network approach for optimized merging of mammographic features. Acad Radiol 2(10):841-850, 1995
2. Lehman CD, Wellman RD, Buist DSM, et al: Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 175(11):1828-1837, 2015. https://doi.org/10.1001/jamainternmed.2015.5231
3. Lee H, Lee EJ, Ham S, et al: Machine learning approach to identify stroke within 4.5 hours. Stroke 51(3):860-866, 2020. https://doi.org/10.1161/STROKEAHA.119.027611
4. Fleuren LM, Klausch TLT, Zwager CL, et al: Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 46(3):383-400, 2020. https://doi.org/10.1007/s00134-019-05872-y
5. Huang Y, Talwar A, Chatterjee S, et al: Application of machine learning in predicting hospital readmissions: A scoping review of the literature. BMC Med Res Methodol 21(1):96, 2021. https://doi.org/10.1186/s12874-021-01284-z
6. Tomašev N, Glorot X, Rae JW, et al: A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 572 (7767):116-119, 2019. https://doi.org/10.1038/s41586-019-1390-1
7. Cao J, Zhang X, Shahinian V, et al: Generalizability of an acute kidney injury prediction model across health systems. Nat Mach Intell 4 (12):1121-1129, 2022. https://doi.org/10.1038/s42256-022-00563-8
8. Gulshan V, Peng L, Coram M, et al: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316(22):2402-2410, 2016. https://doi.org/10.1001/jama.2016.17216
9. Gulshan V, Rajan RP, Widner K, et al: Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmol 137(9):987-993, 2019. https://doi.org/10.1001/jamaophthalmol.2019.2004
10. Esteva A, Kuprel B, Novoa RA, et al: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115-118, 2017. https://doi.org/10.1038/nature21056
11. Liu D, Shin WY, Sprecher E, et al: Machine learning approaches to predicting no-shows in pediatric medical appointment. NPJ Digit Med. 5 (1):1-11, 2022. https://doi.org/10.1038/s41746-022-00594-w
12. Soy Chen, Danielle Bergman, Kelly Miller, et al. Using applied machine learning to predict healthcare utilization based on socioeconomic determinants of care. 2020;26. Accessed February 2, 2023. Available at: https://www.ajmc.com/view/using-applied-machine-learning-to-

predict-healthcare-utilization-based-on-socioeconomic-determinants-of-care
13. Kang J, Schwartz R, Flickinger J, et al: Machine learning approaches for predicting radiation therapy outcomes: A clinician's perspective. Int J Radiat Oncol Biol Phys 93(5):1127-1135, 2015. https://doi.org/10.1016/j.ijrobp.2015.07.2286
14. Kang J, Coates JT, Strawderman RL, et al: Genomics models in radiotherapy: From mechanistic to machine learning. Med Phys 47(5):e203-e217, 2020. https://doi.org/10.1002/mp.13751
15. Hong JC, Eclov NCW, Dalal NH, et al: System for high-intensity evaluation during radiation therapy (SHIELD-RT): A prospective randomized study of machine learning–directed clinical evaluations during radiation and chemoradiation. J Clin Oncol 38(31):3652-3661, 2020. https://doi.org/10.1200/JCO.20.01688
16. Manz CR, Chen J, Liu M, et al: Validation of a machine learning algorithm to predict 180-day mortality for outpatients with cancer. JAMA Oncol 6(11):1723-1730, 2020. https://doi.org/10.1001/jamaoncol.2020.4331
17. Manz CR, Parikh RB, Small DS, et al: Effect of integrating machine learning mortality estimates with behavioral nudges to clinicians on serious illness conversations among patients with cancer: A stepped-wedge cluster randomized clinical trial. JAMA Oncol 2020:e204759. https://doi.org/10.1001/jamaoncol.2020.4759. Published online October 15
18. Sher DJ, Avkshtol V, Moon D, et al: Recurrence and quality-of-life following involved node radiotherapy for head and neck squamous cell carcinoma: Initial results from the phase II INRT-air trial. Int J Radiat Oncol Biol Phys 111(3):e398, 2021. https://doi.org/10.1016/j.ijrobp.2021.07.1155
19. Gensheimer MF, Gupta D, Patel MI, et al: Use of machine learning and lay care coaches to increase advance care planning conversations for patients with metastatic cancer. JCO Oncol Pract 19(2):e176-e184, 2022. https://doi.org/10.1200/OP.22.00128
20. Sher D: DARTBOARD: Novel head & neck cancer trial targets personalized, daily radiation therapy. Dallas: MedBlog, 2022. Published March 2. Accessed February 6, 2023 http://utswmed.org/medblog/head-neck-cancer-clinical-trial/
21. Hong JC, Niedzwiecki D, Palta M, et al: Predicting emergency visits and hospital admissions during radiation and chemoradiation: An internally validated pretreatment machine learning algorithm. JCO Clin Cancer Inform 2:1-11, 2018. https://doi.org/10.1200/CCI.18.00037
22. Taniyama TK, Hashimoto K, Katsumata N, et al: Can oncologists predict survival for patients with progressive disease after standard chemotherapies? Curr Oncol 21(2):84-90, 2014. https://doi.org/10.3747/co.21.1743
23. Kim YJ, Yoon SJ, Suh SY, et al: Performance of clinician prediction of survival in oncology outpatients with advanced cancer. PLoS One 17, 2022(4):e0267467. https://doi.org/10.1371/journal.pone.0267467
24. Temel JS, Greer JA, Muzikansky A, et al: Early palliative care for patients with metastatic non–small-cell lung cancer. N Engl J Med 363(8):733-742, 2010. https://doi.org/10.1056/NEJMoa1000678
25. Gaspar L, Scott C, Rotman M, et al: Recursive partitioning analysis (RPA) of prognostic factors in three Radiation Therapy Oncology Group (RTOG) brain metastases trials. Int J Radiat Oncol Biol Phys 37(4):745-751, 1997. https://doi.org/10.1016/s0360-3016(96)00619-0
26. Sperduto PW, Kased N, Roberge D, et al: Summary report on the graded prognostic assessment: An accurate and facile diagnosis-specific tool to estimate survival for patients with brain metastases. J Clin Oncol 30(4):419-425, 2012. https://doi.org/10.1200/JCO.2011.38.0527
27. Sperduto PW, Yang TJ, Beal K, et al: Estimating survival in patients with lung cancer and brain metastases: An update of the graded prognostic assessment for lung cancer using molecular markers (Lung-molGPA). JAMA Oncol 3(6):827-831, 2017. https://doi.org/10.1001/jamaoncol.2016.3834
28. Bell EH, Pugh SL, McElroy JP, et al: Molecular-based recursive partitioning analysis model for glioblastoma in the temozolomide era: A correlative analysis based on NRG oncology RTOG 0525. JAMA Oncol 3(6):784-792, 2017. https://doi.org/10.1001/jamaoncol.2016.6020
29. Parikh RB, Manz C, Chivers C, et al: Machine learning approaches to predict 6-month mortality among patients with cancer. JAMA Netw Open 2, 2019(10). https://doi.org/10.1001/jamanetworkopen.2019.15997. e1915997-e1915997
30. Gensheimer MF, Henry AS, Wood DJ, et al: Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. J Natl Cancer Inst 111(6):568-574, 2019. https://doi.org/10.1093/jnci/djy178
31. Gensheimer MF, Aggarwal S, Benson KRK, et al: Automated model versus treating physician for predicting survival time of patients with metastatic cancer. J Am Med Inform Assoc 28(6):1108-1116, 2021. https://doi.org/10.1093/jamia/ocaa290
32. Chera BS, Amdur RJ, Green R, et al: Phase II trial of de-intensified chemoradiotherapy for human papillomavirus–associated oropharyngeal squamous cell carcinoma. JCO 37(29):2661-2669, 2019. https://doi.org/10.1200/JCO.19.01007
33. Ferris RL, Flamand Y, Weinstein GS, et al: Phase II randomized trial of transoral surgery and low-dose intensity modulated radiation therapy in resectable p16+ locally advanced oropharynx cancer: An ECOG-ACRIN Cancer Research Group Trial (E3311). J Clin Oncol 40(2):138-149, 2022. https://doi.org/10.1200/JCO.21.01752
34. Yom SS, Torres-Saavedra P, Caudell JJ, et al: Reduced-dose radiation therapy for HPV-associated oropharyngeal carcinoma (NRG oncology HN002). JCO 39(9):956-965, 2021. https://doi.org/10.1200/JCO.20.03128
35. National Cancer Institute (NCI). *A Randomized Phase II/III Trial of De-Intensified Radiation Therapy for Patients With Early-Stage, P16-Positive, Non-Smoking Associated Oropharyngeal Cancer*. 2023. Accessed January 5, 2023. Available at: https://clinicaltrials.gov/ct2/show/NCT03952585
36. Bernier J, Cooper JS, Pajak TF, et al: Defining risk levels in locally advanced head and neck cancers: A comparative analysis of concurrent postoperative radiation plus chemotherapy trials of the EORTC (#22931) and RTOG (# 9501). Head Neck 27(10):843-850, 2005. https://doi.org/10.1002/hed.20279
37. Almulla A, Noel CW, Lu L, et al: Radiologic-pathologic correlation of extranodal extension in patients with squamous cell carcinoma of the oral cavity: Implications for future editions of the TNM classification. Int J Radiat Oncol Biol Phys 102(4):698-708, 2018. https://doi.org/10.1016/j.ijrobp.2018.05.020
38. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 521(7553):436-444, 2015. https://doi.org/10.1038/nature14539
39. Kann BH, Aneja S, Loganadane GV, et al: Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks. Sci Rep 8(1):14036, 2018. https://doi.org/10.1038/s41598-018-32441-y
40. Kann BH, Hicks DF, Payabvash S, et al: Multi-institutional validation of deep learning for pretreatment identification of extranodal extension in head and neck squamous cell carcinoma. JCO 38(12):1304-1311, 2019. https://doi.org/10.1200/JCO.19.02031
41. O'Sullivan B, Huang SH, de Almeida JR, et al: Alpha test of intelligent machine learning in staging head and neck cancer. J Clin Oncol 38(12):1255-1257, 2020. https://doi.org/10.1200/JCO.19.03309
42. Simon AB, Vitzthum LK, Mell LK: Challenge of directly comparing imaging-based diagnoses made by machine learning algorithms with those made by human clinicians. JCO 38(16):1868-1869, 2020. https://doi.org/10.1200/JCO.19.03350
43. Kann BH, Payabvash S, Aneja S: Reply to A.B. Simon et al. JCO 38(16):1869-1870, 2020. https://doi.org/10.1200/JCO.20.00402
44. Kann BH, Likitlersuang J, Ye Z, et al: Screening for extranodal extension with deep learning: Evaluation in ECOG-ACRIN E3311, a randomized de-escalation trial for HPV-associated oropharyngeal carcinoma. Int J Radiat Oncol Biol Phys 114(3):S26-S27, 2022. https://doi.org/10.1016/j.ijrobp.2022.07.379
45. Sher DJ, Pham NL, Shah JL, et al: Prospective phase 2 study of radiation therapy dose and volume de-escalation for elective neck treatment of oropharyngeal and laryngeal cancer. Int J Radiat Oncol Biol Phys 109(4):932-940, 2021. https://doi.org/10.1016/j.ijrobp.2020.09.063
46. Chen L, Zhou Z, Sher D, et al: Combining many-objective radiomics and 3D convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer. Phys Med Biol 64, 2019(7):075011. https://doi.org/10.1088/1361-6560/ab083a

47. Sher D. *INRT- AIR: A Prospective Phase II Study of Involved Nodal Radiation Therapy Using Artificial Intelligence-Based Radiomics for Head and Neck Squamous Cell Carcinoma*. clinicaltrials.gov; 2022. Available at: https://clinicaltrials.gov/ct2/show/NCT03953976. Accessed February 2, 2023
48. Esteva A, Feng J, van der Wal D, et al: Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. NPJ Digit Med. 5(1):1-8, 2022. https://doi.org/10.1038/s41746-022-00613-w
49. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: 2016:770-778.
50. Heukelom J, Fuller CD: Head and neck cancer adaptive radiation therapy (ART): Conceptual considerations for the informed clinician. Semin Radiat Oncol 29(3):258-273, 2019. https://doi.org/10.1016/j.semradonc.2019.02.008
51. Archambault Y, Boylan C, Bullock D, et al: Medical physics international. Med Phys Int 8(2):77-86, 2020
52. Varian A Siemens Healthiness Company Varian clinical research. Accessed February 5, 2023. Available at: https://medicalaffairs.varian.com/clinical-research
53. Mandrekar JN: Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 5(9):1315-1316, 2010. https://doi.org/10.1097/JTO.0b013e3181ec173d
54. Christensen E: Methodology of diagnostic tests in hepatology. Ann Hepatol 8(3):177-183, 2009. https://doi.org/10.1016/S1665-2681(19)31763-6
55. Mayo DG, Hand D: Statistical significance and its critics: Practicing damaging science, or damaging scientific practice? Synthese 200 (3):220, 2022. https://doi.org/10.1007/s11229-022-03692-0
56. Azzolina D, Lorenzoni G, Bressan S, et al: Handling poor accrual in pediatric trials: A simulation study using a Bayesian approach. Int J Environ Res Public Health 18(4):2095, 2021. https://doi.org/10.3390/ijerph18042095
57. Liao Z, Lee JJ, Komaki R, et al: Bayesian adaptive randomization trial of passive scattering proton therapy and intensity-modulated photon radiotherapy for locally advanced non-small-cell lung cancer. J Clin Oncol 36 (18):1813-1822, 2018. https://doi.org/10.1200/JCO.2017.74.0720
58. Lin SH, Hobbs BP, Verma V, et al: Randomized phase IIB trial of proton beam therapy versus intensity-modulated radiation therapy for locally advanced esophageal cancer. J Clin Oncol 38(14):1569-1579, 2020. https://doi.org/10.1200/JCO.19.02503
59. Spiegelhalter DJ, Freedman LS, Parmar MKB: Bayesian approaches to randomized trials. J R Stat Soc 157(3):357-387, 1994
60. Mayo DG: Don't throw out the error control baby with the bad statistics bathwater: A commentary. Am Stat 70:129-133, 2016
61. Breiman L: Statistical Modeling: The two cultures (with comments and a rejoinder by the author). Statist Sci 16(3):199-231, 2001. https://doi.org/10.1214/ss/1009213726
62. Molnar C. *Interpretable machine learning*. 2019. Accessed December 24, 2019. Available at: https://christophm.github.io/interpretable-ml-book/
63. Idnay B, Dreisbach C, Weng C, et al: A systematic review on natural language processing systems for eligibility prescreening in clinical research. J Am Med Inform Assoc 29(1):197-206, 2021. https://doi.org/10.1093/jamia/ocab228
64. Liu R, Rizzo S, Whipple S, et al: Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature 592(7855):629-633, 2021. https://doi.org/10.1038/s41586-021-03430-5
65. Lazer D, Kennedy R, King G, et al: The parable of google flu: Traps in big data analysis. Science 343(6176):1203-1205, 2014. https://doi.org/10.1126/science.1248506
66. Park AL: Injustice ex Machina: Predictive algorithms in criminal sentencing. UCLA Law Review. Published online February 19. Accessed February 19, 2023. Available at: https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/
67. Conceptual challenges for interpretable machine learning | SpringerLink. Accessed February 19, 2023. Available at: https://link.springer.com/article/10.1007/s11229-022-03485-5
68. Chowdhry AK, Gondi V, Pugh SL: Missing data in clinical studies. Int J Radiat Oncol Biol Phys 110(5):1267-1271, 2021. https://doi.org/10.1016/j.ijrobp.2021.02.042
69. Hernan MA, Robins JM: Causal Inference: What If. Boca Raton: CRC Press, 2023.
70. Beede E, Baylor E, Hersch F, et al: A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, ACM, 1-12, 2020. https://doi.org/10.1145/3313831.3376718