Surgery 168 (2020) 253–266

Contents lists available at ScienceDirect

Surgery

journal homepage: www.elsevier.com/locate/surg

Decision analysis and reinforcement learning in surgical decision-making

Tyler J. Loftus, MD[a], Amanda C. Filiberto, MD[a], Yanjun Li, MS[b], Jeremy Balch, MD[a], Allyson C. Cook, MD[c], Patrick J. Tighe, MD[d], Philip A. Efron, MD[a], Gilbert R. Upchurch Jr., MD[a], Parisa Rashidi, PhD[e,f], Xiaolin Li, PhD[b], Azra Bihorac, MD[c,f,*]

[a] Department of Surgery, University of Florida Health, Gainesville, FL
[b] NSF Center for Big Learning, University of Florida, Gainesville, FL
[c] Department of Medicine, University of California, San Francisco, CA
[d] Departments of Anesthesiology, Orthopedics, and Information Systems/Operations Management, University of Florida Health, Gainesville, FL
[e] Departments of Biomedical Engineering, Computer and Information Science and Engineering, and Electrical and Computer Engineering, University of Florida, Gainesville, FL
[f] Precision and Intelligence in Medicine, Department of Medicine, University of Florida Health, Gainesville, FL

A R T I C L E I N F O

Article history:
Available online 13 June 2020

A B S T R A C T

Background: Surgical patients incur preventable harm from cognitive and judgment errors made under time constraints and uncertainty regarding patients' diagnoses and predicted response to treatment. Decision analysis and techniques of reinforcement learning theoretically can mitigate these challenges but are poorly understood and rarely used clinically. This review seeks to promote an understanding of decision analysis and reinforcement learning by describing their use in the context of surgical decision-making.
Methods: Cochrane, EMBASE, and PubMed databases were searched from their inception to June 2019. Included were 41 articles about cognitive and diagnostic errors, decision-making, decision analysis, and machine-learning. The articles were assimilated into relevant categories according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews guidelines.
Results: Requirements for time-consuming manual data entry and crude representations of individual patients and clinical context compromise many traditional decision-support tools. Decision analysis methods for calculating probability thresholds can inform population-based recommendations that jointly consider risks, benefits, costs, and patient values but lack precision for individual patient-centered decisions. Reinforcement learning, a machine-learning method that mimics human learning, can use a large set of patient-specific input data to identify actions yielding the greatest probability of achieving a goal. This methodology follows a sequence of events with uncertain conditions, offering potential advantages for personalized, patient-centered decision-making. Clinical application would require secure integration of multiple data sources and attention to ethical considerations regarding liability for errors and individual patient preferences.
Conclusion: Traditional decision-support tools are ill-equipped to accommodate time constraints and uncertainty regarding diagnoses and the predicted response to treatment, both of which often impair surgical decision-making. Decision analysis and reinforcement learning have the potential to play complementary roles in delivering high-value surgical care through sound judgment and optimal decision-making.

Introduction

Every day, patients and physicians must decide which diagnostic and therapeutic interventions should be performed or deferred. Although hundreds or thousands of interventions may yield more benefit than harm, limitations of time and resources mandate

* Corresponding author: Azra Bihorac, MD, MS, Department of Medicine, University of Florida Health, Division of Nephrology, Hypertension, and Renal Transplantation, PO Box 100224, Gainesville, FL 32610-0224.
E-mail address: abihorac@ufl.edu (A. Bihorac).

https://doi.org/10.1016/j.surg.2020.04.049
0039-6060/

that only the most advantageous interventions be performed. This approach to resource use is often misused or ignored in the United States, where doctors and hospitals may tend to overtreat the insured and undertreat the uninsured.[1] More importantly, decisions regarding interventions affect mortality, morbidity, and quality of life for patients and their caregivers.

Ideally, clinical reasoning incorporates rigorous medical training, clinical intuition, critical thinking, evidence-based medicine, and a robust process of shared decision-making among physicians, patients, and their caregivers. Unfortunately, decisions often transpire under time constraints and conditions of uncertainty regarding an individual patient's diagnoses and predicted response to treatment. Time constraints may be imposed by acute diseases that require urgent diagnosis and treatment, or by busy clinical schedules that restrict time for gathering information and deliberating. Uncertainty may be imposed by a lack of provider knowledge, the unavailability of patient data, such as outside hospital records or diagnostic tests, or the absence of high-level evidence to guide important management decisions. Under such time constraints and uncertainty, clinicians may rely instead on cognitive shortcuts and snap judgments using pattern recognition and intuition.[2,3] Cognitive shortcuts without deliberation can lead to bias or predictable and systematic cognitive errors.[4,5] Cognitive and judgment errors are a leading cause of misdiagnosis, and physicians are often blind to them unless feedback is provided by postmortem examinations, of which 10% to 15% reveal major diagnostic errors.[6–8] Cognitive and judgment errors are especially harmful in surgical decision-making, in which high-stakes decisions can markedly affect clinical outcomes.[9] In a survey of 7,905 members of the American College of Surgeons, lapses in judgment were the most common cause of major medical errors.[10]

Decision-support tools are intended to mitigate these errors. Unfortunately, they often require time-consuming, manual data entry and are designed for nonspecific, generalized application to any patient with a certain disease or condition, and so they lack precision for the unique pathophysiology and clinical context of individual patients.[11] Not surprising, most of these decision-support tools have not achieved widespread clinical adoption.[12] Surgeons need better decision-support tools. Decision analysis methods and technologies of reinforcement learning can generate population-based recommendations and augment decision-making for individual patients. Unfortunately, many clinicians are unfamiliar with them and the applications to surgery are sparse. Among many promising methods for improving patient-centered decision support,[13,14] this review features reinforcement learning because it most closely mimics human learning and offers specific recommendations for discrete actions rather than predicted probabilities that only indirectly support decisions. Predictive analytic risk assessments are useful when risk is unexpectedly very low or very high, but most patients have intermediate risk. This review describes decision analysis and reinforcement learning in the context of clinical surgical decision-making.

Methods

Cochrane, EMBASE, and PubMed databases were searched from their inception to June 2019. Supplemental Digital Content 1 lists article search parameters and objectives. Articles were excluded if they were not published in English or were not primary literature or a review article. Articles were selected for inclusion after we read the abstracts and full texts to assess topical relevance, methodologic strength, and novel or meritorious contribution to existing literature. Articles of interest cited by other articles identified in the initial search were reviewed using the same inclusion criteria. We included 41 articles and assimilated them into relevant categories (Table) according to guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews. Supplemental Digital Content 2 lists Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews criteria. The determination to review decision analysis and reinforcement learning methods was made before performing the literature search. Topic subcategories were chosen after performing the literature review by favoring themes that emerged from the literature. Decision-making concepts and theories were described in the context of surgical decision-making scenarios. The assimilation process was limited by heterogeneity among topics and reporting practices that precluded the performance of a systematic review and meta-analysis. The 41 articles included addressed the topics of decision-making ($n = 13$), decision analysis ($n = 13$), and machine-learning ($n = 15$).

Observations

Patient-centered decision-making

Shared decision-making that is truly effective improves patient satisfaction and compliance and may decrease costs from unnecessary interventions.[15,16] Ethically, patient-centered decision-making should be a fundamental principle governing a health care system that values patient autonomy.[17] But clinicians often ignore patient values. Patients, caregivers, and providers frequently misunderstand one another and their goals of care.[16,18] These misunderstandings are compounded not only when patients and caregivers with limited health literacy make complex medical decisions, but also when clinicians fail to recognize inadequate decision-making capacity. Bertrand et al[19] assessed the decision-making capacity of 206 patients in an intensive care unit (ICU), using 2 methods: a mini-mental status examination and the opinion of attending physicians, nurses, and residents. Clinicians believed 45% of the population had decision-making capacity, but only 17% of the patients had the capacity according to the criteria of the mini-mental status examination. In a systematic review of 32 articles including 13,176 patients and surgeons, only 36% of all patient–surgeon interactions represented shared decision-making.[20] Surgeons are often unknowingly blind to this phenomenon, and 1 in 7 surgical patients report decisional regret.[20,21] After establishing rapport and decision-making capacity, surgeons should ask patients about their goals of care and values. These findings suggest that patient assessments often omit this step.

Research that should rely on patient preferences often omits these patient preferences. Noninferiority trials measure a trade-off between losing the established efficacy of a standard treatment and some possible benefit of a new therapy. If investigators weigh risk-benefit trade-offs differently than patients, the new therapy may be designated noninferior and achieve clinical adoption before clinicians realize that patients actually preferred the standard therapy. Acuna et al[22] cite the American College of Surgeons Oncology Group Z0011 trial as an example. Patients who did not undergo completion of axillary lymph node dissection had 45% lesser rates of surgical complications and 13% lesser rates of lymphedema at 1-year follow-up, and the noninferiority margin for overall survival set by investigators was 1.3, or 6%.[23] But some patients may not accept a 6% decreased overall survival in exchange for fewer complications.

Many prediction models and decision-support tools ignore patient values. Each year, more than 1,000 published articles feature independent risk factors or independent predictors in their title or abstract.[12] Most decision-support tools described in these articles have flaws that preclude their widespread clinical adoption (eg, the risk factors or predictors are not widely available or used by

Table I
Summary of included studies

Authors	Study topic	Study design	Population	Sample size	Major findings pertinent to this review	Sources of funding, conflicts of interest
Dijksterhuis et al[3]	Decision-making	Observational	Consumers of simple and complex products	75	Conscious deliberation can lead to less satisfactory decision-making than "deliberation-without-attention" choices, especially for complex purchases.	Netherlands Organization for Scientific Research
Wolf et al[5]	Decision-making	Prospective	First-year house officers	89	Bayesian inference can provide a useful complement to clinical judgment, although it is rarely used by first-year house officers.	Spencer Foundation
Graber et al[6]	Decision-making	Retrospective	Cases of suspected diagnostic error	100	System-related factors and cognitive biases result in diagnostic error.	National Patient Safety Foundation
Kirch and Schafii[7]	Decision-making	Retrospective	Autopsy reports from 1959, 1969, 1979, and 1989	400	Misdiagnosis rates of were approximately 10% across all decades.	None reported
Sonderegger-Iseli et al[8]	Decision-making	Retrospective	Autopsy reports from 1972, 1982, and 1992	300	Frequency of major discrepancy between clinical diagnosis and necropsy findings in 1992 was 14%.	None reported
Healey et al[9]	Decision-making	Prospective	Surgical inpatients	4,658	Half of all adverse events were attributable to provider error. Diagnostic and judgment errors were the second most common cause of preventable harm.	None reported
Shanafelt et al[10]	Decision-making	Cross-sectional	Members of the American College of Surgeons	7,905	Surgeons who attributed error to individual, rather than system-level, factors (70%). Surgeons who reported making a major medical error in the last 3 months (9%), and lapses in judgment were the most common cause (32%).	None reported
Leeds et al[11]	Decision-making	Cross-sectional	Surgical trainees at 4 institutions	124	A total of 26% of surgical trainees report using validated, contemporary risk communication frameworks. Barriers to use included lack of electronic and clinical workflow integration.	NIH, NCI, ASCRS, AHRQ
Brotman et al[12]	Decision-making	MEDLINE search	Articles with "independent risk factor" or "independent predictor" in their title or abstract	—	Each year, more than 1,000 articles are published investigating "independent risk factors" or "independent predictors."	None reported
Legare et al[18]	Decision-making	Systematic review	Articles about implementing shared decision-making practices	38	Barriers to shared decision-making included time constraints and lack of applicability to patient and clinical context. Facilitators were provider motivation and positive impact on clinical process and patient outcomes.	Canada Research Chair in Implementation of Shared Decision-Making
Bertrand et al[19]	Decision-making	Cross-sectional	ICU patients and their providers	419	Decision-making capacity was overestimated by providers, largely because of inappropriate conflation of consciousness and decision-making capacity.	Gabriel Montpied Teaching Hospital, Pfizer, Fisher & Paykel, Gilead, Jazz Pharma, Baxter, Astellas, Alexion
de Mik et al[20]	Decision-making	Systematic review	Literature on shared decision-making during surgical consultations	32	Only 36% of all patients and surgeons perceived the consultation as shared decision-making. Surgeons were more likely to perceive that interactions represented shared decision-making.	AMC Foundation
Wilson et al[21]	Decision-making	Systematic review	Literature on self-reported decisional regret	79	Patients who reported regret (14.4%), most often associated with the type of surgery, disease-specific quality of life, and shared decision-making	None reported
Guyatt et al[31]	Decision analysis	Review	—	—	Demonstrates integration of patient values with decision analysis for risks of stroke and hemorrhage when prescribing antiplatelet therapy for patients with atrial fibrillation.	None reported

(continued on next page)

Table I (*continued*)

Authors	Study topic	Study design	Population	Sample size	Major findings pertinent to this review	Sources of funding, conflicts of interest
Pauker and Kassirer[33]	Decision analysis	Review	—	—	Establishment of "testing" threshold and a "test-treatment" threshold to guide decision-making on treatment and diagnostic testing.	NIGMS, NLM, NIH
O'Brien et al[39]	Decision analysis	Review	—	—	Description of how to perform cost-adjusted value measures like quality-adjusted life years.	None reported
Djulbegovic et al[36]	Decision analysis	Review	—	—	Applies patient values to decision analysis for DVT prophylaxis.	None reported
Vickers and Elkin[37]	Decision analysis	Review	—	—	Description of how decision curve analysis can be used to evaluate diagnostic and prognostic strategies .	NCI SPORE
Gage et al[40]	Decision analysis	Decision analysis	Patients with nonvalvular atrial fibrillation	—	Warfarin is cost-effective for patients with nonvalvular atrial fibrillation and one additional stroke risk factor. In patients without such risk factors, this benefit was lost.	None reported
Robbins et al[41]	Decision analysis	Cost-benefit analysis	Infants treated with RSV-IG from 3 RCT	1,108	Demonstrates the use of number-needed-to-treat principles in determining RSV-IG treatment for specific infant populations.	None reported
Komorowski et al[45]	Machine learning	Retrospective review	Patients admitted to ICU with sepsis	96,156	A reinforcement learning model recommended intravenous fluid and vasopressor strategies, mortality was lowest when decisions made by clinicians matched recommendations from the reinforcement learning model.	NIHR, EPSRC, Orion Corp, Amomed Pharma, Ferring Pharma, Tenax Therapeutics, Baxter, Bristol-Myers Squibb, GSK, HCA International, Philips Health care, Fresenius-KABI
Silver et al[46]	Machine learning	Observational	Go board game	—	A deep reinforcement learning model trained by human expert moves and self-play provided high-fidelity victories against previous Go algorithms and human experts.	Google, Google DeepMind
Mnih et al[47]	Machine learning	Observational	Atari games	—	Demonstrated development of a deep Q network to incorporate highly dimensional sensory inputs and actions to optimize machine performance in Atari video games.	Google, Google DeepMind
Shickel et al[49]	Machine learning	Retrospective review	ICU admissions	79,701	A deep model fed with SOFA variables predicted in-hospital mortality with greater accuracy than the traditional SOFA score (AUC 0.90 vs 0.85).	NIGMS, NSF, University of Florida CTSI, NCATS, J Crayton Pruitt Family Department of Biomedical Engineering, NVIDIA
Sundaram et al[50]	Machine learning	Observational	Actual and simulated animal and human genomes	—	A deep neural network identified pathogenic mutations for rare diseases, with 88% accuracy and discovered candidate genes for intellectual disability.	Health Innovation Challenge Fund, Wellcome Sanger Institute, NIHR, NIGMS, NSF
Li et al[51]	Machine learning	Observational	Actual and simulated protein sequences	—	A deep neural network predicted protein properties by learning from protein sequences, with no supervision or domain knowledge.	NSF, NIH, Industrial Members of NSF Center for Big Learning
Rajpurkar et al[52]	Machine learning	Retrospective review	Chest radiographs	420	A deep learning algorithm had equivalent performance to board-certified radiologists in 10/14 pathologies, superior performance in 1/14, and inferior performance in 3/14.	Stanford AIMI Center, whiterabbit.ai, nines.ai, Nuance communications, Radiological Society of North America, Phillips Healthcare, GE Healthcare
Davoudi et al[53]	Machine learning	Observational	Surgical ICU patients	22	Autonomous collection of granular patient and environmental data can identify contributors to delirium.	NSF CAREER, NIH/NIGMS, NIH, NIH/NIBIB
Hashimoto et al[54]	Machine learning	Observational	Laparoscopic sleeve gastrectomy cases	88	The use of computer vision and deep neural networks can identify quantitative steps in operative procedures.	NIH, Olympus Corporation, Toyota Research Institute, Verily Life Sciences, Johnson & Johnson Institute, Gerson Lehrman Group
Silver et al[56]	Machine learning	Observational	Go board game		A deep reinforcement learning model trained by self-play with no human input consistently defeated a version that used human input.	Google, Google DeepMind

Table I (*continued*)

Authors	Study topic	Study design	Population	Sample size	Major findings pertinent to this review	Sources of funding, conflicts of interest
Pineau et al[65]	Machine learning	Review	Experimental epilepsy models	—	Reinforcement learning paradigms of EEG measurements can be used to optimize electrostimulations patterns in the treatment of epilepsy.	NSERC, CIHR
Van Calster et al[66]	Decision analysis	Systematic review	Men undergoing prostate biopsy	3,616	Demonstrates the use of decision curve analysis to identify a range of clinically reasonable risk thresholds for prostate biopsy.	Research Foundation— Flanders
Tinetti et al[67]	Decision analysis	Review	Disease-specific guidelines	—	Adherence to disease-specific guidelines in patients with multiple chronic conditions may result in clinical harm.	NIA, VA HSR&D, Merck, AFAR
Boyd et al[68]	Decision analysis	Observational	Clinical practice guidelines for Medicare beneficiaries	—	Adherence to clinical practice guidelines for disease-specific entities may result in suboptimal care for elderly patients with multiple comorbidities.	NIH, NIA, HRSA, Roger C. Lipitz Center for Integrated Health Care, Partnership for Solutions
Che et al[70]	Machine learning	Retrospective review	Critically ill children	398	Deep learning models often lack interpretability. Shallow models and knowledge-distillation approaches can clarify underlying processes for clinicians.	NSF, USC Coulter Translational Research Program
Gal[71]	Machine learning	Dissertation	—	—	Using a softmax function to map machine learning output layer, network activations may overestimate model confidence that its outputs are accurate.	Google AI, Qualcomm
Guo et al[72]	Machine learning	Review	Machine learning models in published literature	—	Many descriptions of machine learning models do not incorporate and report calibration.	NSF, Bill and Melinda Gates Foundation, Office of Naval Research
Vergouwe et al[73]	Decision analysis	Observational	Moderate or severe brain injury patients	1,118	Creating benchmark values that incorporate distributions of patient characteristics can improve external validity of prediction models.	Netherlands Organization for Scientific Research, NIH
Van Calster et al[74]	Decision analysis	Observational	Decision analysis models	—	Miscalibration of a model (overestimation, underestimation, overfitting, and underfitting) to a baseline event rate reduces net benefit and can impair clinical decision-making.	Research Foundation-Flanders
Goldstein et al[75]	Machine learning	Retrospective review	Hemodialysis patients	18,846	Comparing summary statistics, machine learning methods, functional data analysis, and joint models revealed that complex approaches using highly dimensional EHR data may impair mortality predictions.	NIDDK
McGlynn et al[76]	Decision analysis	Cross-sectional study	Randomly selected patients from 12 US metropolitan areas	13,275	Only slightly more than half of all patients surveyed received care recommended by clinical practice guidelines.	Robert Wood Johnson Foundation, VA HSR&D

AFAR, American Federation for Aging Research; *AHRQ*, Agency for Healthcare Research and Quality; *AI*, artificial intelligence; *AIMI*, Artificial Intelligence in Medicine and Imaging; *ASCRS*, American Society of Colon and Rectal Surgeons; *CIHR*, Canadian Institute of Health Research; *EPSRC*, Engineering and Physical Sciences Research Council; *HRSA*, Health Resources and Services Administration; *NCI*, National Cancer Institute; *NIA*, National Institute on Aging; *NIBIB*, National Institute on Biomedical Imaging and Bioengineering; *NIGMS*, National Institute of General Medical Sciences; *NIH*, National Institutes of Health; *NIHR*, National Institute for Health Research; *NLM*, National Library of Medicine; *NSERC*, Natural Sciences and Engineering Research Council; *NSF*, National Science Foundation; *SOFA*, Sequential Organ Failure Assessment; USC, University of Southern California; *VA HSR&D*, Veterans Affairs Health Services Research and Development.

clinicians, predictive performance is weak, or the findings are not validated in a separate study population to ensure generalizability), but even the successful tools often do not incorporate patient values. The CHA2DS2-VASc score is a clinical classification scheme that uses 7 ordinal and binary variables to estimate annualized risk of stroke among patients with atrial fibrillation and makes clinically useful recommendations regarding antiplatelet and anticoagulation therapy, earning support from the European Society of Cardiology, American College of Cardiology, and the American Heart Association.[24] CHA2DS2-VASc makes assumptions about patient preferences for outcomes like stroke and hemorrhage, which may skew decisions regarding antiplatelet and anticoagulation therapy for any individual patient, as discussed in the "Patient values" section later in this review..

Decision analysis

Clinical and translational research and evidence-based medicine define best practices for managing disease and for promoting health by measuring and evaluating the risks and benefits of diagnostic and therapeutic interventions. Clinical application requires the additional step of considering these risks and benefits

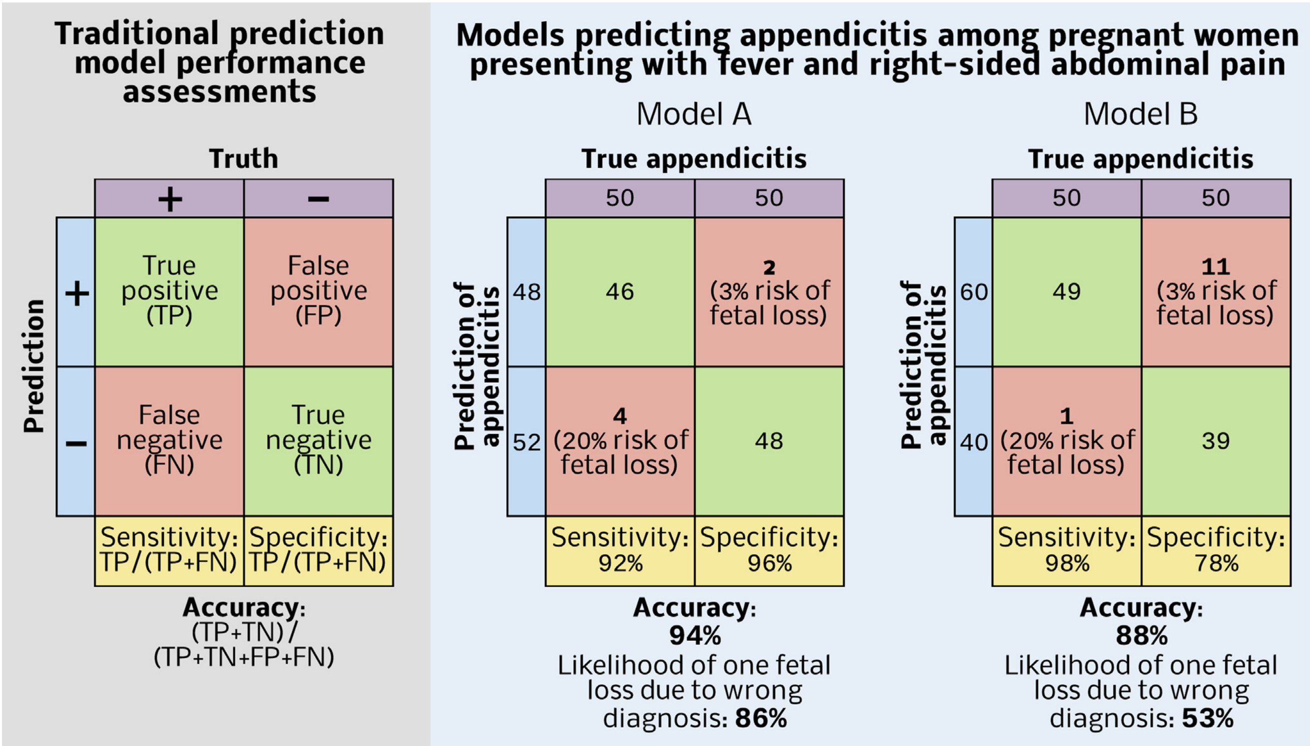

Fig 1. Optimizing the accuracy of the prediction model may not optimize clinical utility. Model A has greater accuracy, but, if a pregnant woman presenting with fever and right-sided abdominal pain wishes to avoid fetal demise because of a wrong diagnosis, then model B is favorable.

alongside patient values and financial costs. Methods of decision analysis accomplish this step by weighing the risks and benefits by patient values and by incorporating costs to quantify the value of care, thereby facilitating the optimal use of resources across health care systems. This process produces probability thresholds that inform guidelines and recommendations for diagnostic and therapeutic interventions across populations.

Evaluating model utility

The diagnostic performance and clinical utility of a test or model are complementary but separate considerations. A magnetic resonance image of the chest may have excellent diagnostic performance in identifying traumatic thoracic injuries, but obtaining a chest magnetic resonance image for an unstable patient with penetrating chest trauma could harm the patient by delaying operative exploration, thereby yielding negative clinical utility. The techniques of decision analysis compare directly the overall clinical utility of diagnostic tests or prediction models based on risks, benefits, costs, and patient values. This offers a major advantage over the common practice of comparing tests and models by discrimination or accuracy alone. For example, the diagnosis of appendicitis among pregnant women is challenging. Several other conditions mimic appendicitis. cephalad displacement of the appendix alters the clinical presentation, and teratogenic radiation effects preclude the routine use of computed tomography. A missed diagnosis with progression to complicated appendicitis is associated with increased risk for fetal loss relative to the risk of nontherapeutic laparotomy, 20% vs 3% in one study.[25] Therefore, in predicting appendicitis among pregnant women, false-negative results are more harmful than false-positives results.

Consider 2 models predicting appendicitis among 100 pregnant women presenting with fever and right-sided abdominal pain, of whom 50 actually have appendicitis (Fig 1). Model A has much greater specificity and slightly less sensitivity than model B. Accuracy assigns equal weight to sensitivity and specificity, therefore model A is more accurate. The likelihood of 1 fetal loss attributed to a wrong diagnosis applying model A 100 times is $(4 \times 0.20) + (2 \times 0.03) = 0.86$. The likelihood of fetal loss, using model B, is $(1 \times 0.20) + (11 \times 0.03) = 0.53$. If a woman wishes to avoid fetal demise attributable to a wrong diagnosis, model B has greater utility, despite its lesser accuracy. In such cases, metrics, like the number needed to treat or harm, are useful.

Number needed to treat or to harm

The number needed to treat (NNT)—the number of patients that must undergo an intervention to avoid one adverse event—adjusts for prevalence by incorporating baseline risk without an intervention and the risk reduction associated with the intervention. The importance of adjusting for prevalence is illustrated by application of Bayesian probability to mammographic detection of breast cancer.[26] A group of physicians was presented with the following 3 statistics: A 40-year-old woman undergoing screening mammography has a 1% chance of having breast cancer. If she has breast cancer, the probability of a positive mammography is 80%. If she does not have breast cancer, the probability of a positive mammography is 9.6%. Most physicians in this study estimated that this 40-year-old woman with a positive screening mammogram had a 70% to 80% probability of actually having breast cancer, approximately 1 order of magnitude greater than the actual probability of 7.8%.

NNT is the reciprocal of absolute risk reduction, or the raw difference in risk of an adverse event between 2 options. Consider an uncomplicated, intra-abdominal infection for which management options include antibiotics alone or surgical source control. If the risk of disease progression and septic shock while treating with antibiotics alone is 7% and the risk or progression and shock after a surgical source control procedure is 2%, then the number needed to treat with surgery to avoid 1 case of septic shock is $1/(0.07 - 0.02) = 20$ patients. NNT does not account for adverse events attributable to the intervention itself, manifest as number needed to harm, or the number of patients that must undergo an intervention to produce 1 adverse event, calculated as the reciprocal of the raw difference in harm. If the risk of allergy or untoward effect from antibiotics is 4% and the risk of postoperative complications is 8%, then the number needed to harm with surgery is $1/ (0.08 - 0.04) = 25$. When NNT = 20 and number needed to harm = 25, surgery is advantageous when assuming equal weight for postoperative complications, medication side effects, and progression to septic shock. Patients and surgeons may not agree with these assumptions. Incorporation of relative value addresses this problem.

Patient values

Probability thresholds incorporate patient values by calculating relative values of risks and benefits attributable to the intervention and its alternatives. Published literature can produce relative values. The CHA2DS2-VASc score makes assumptions about patient values regarding stroke and hemorrhage when recommending antiplatelet and anticoagulation therapy for patients with atrial fibrillation. In 4 studies investigating patient preferences and quality of life, patients appear to consider 1 stroke equivalent to 5 episodes of serious gastrointestinal bleeding.[27–30] Considering this ratio within a decision analysis framework, the relative value of serious bleeding relative to stroke is 0.744. The relative value of minor bleeding relative to stroke is 0.014.[31] Applied to known frequencies of major and minor bleeding events among anticoagulated patients, the threshold NNT is 152. Among elderly patients with a history of stroke, diabetes, and hypertension, anticoagulation decreases 1-year stroke risk from 8.1% to 2.6%, such that the $\text{NNT}=1/(0.081 - 0.026) = 18$.[32] The NNT in this subgroup is well below the threshold NNT, and therefore, this subgroup should receive anticoagulation therapy.

This calculation used five well-designed studies to derive and apply relative values.[27–30,32] Similar data are often unavailable for surgical diseases, and especially for rare ones. In addition, this method calculates thresholds for aggregate patient populations. A patient who declines allogenic blood transfusions may consider stroke and serious gastrointestinal bleeding to be equally harmful, generating a different probability threshold than the general population.

Decision trees and curves

Decision tree analysis uses predicted risks, benefits, and relative values of possible outcomes to calculate probability thresholds.[33] Each patient has a probability p that the disease is present. If p is near 1, a diagnostic or therapeutic intervention targeting that disease is likely useful; in contrast, if p is near 0, the intervention is likely useless. Between 0 and 1, there is a probability threshold p_t where the predicted utilities of performing and deferring the intervention are equal. Decision trees are the foundation for some machine-learning methods. Random Forests use a multitude of decision trees, as the name implies. This review considers decision trees separately from the machine-learning techniques that employ decision trees.

Consider a patient who presents with postprandial epigastric pain (Fig 2). Whether symptoms are attributable to biliary dyskinesia or another process (eg, gastritis, pancreatitis) is unclear. Approximately 60% to 90% of all adults with similar presentations will have improvement or resolution of these symptoms after cholecystectomy, with a lesser likelihood of a benefit for patients with atypical symptoms and no gallstones.[34,35] This thought experiment assumes 75% probability that symptoms are attributable to biliary dyskinesia and will resolve after cholecystectomy. Assume that the value of surgery when disease is present and the value of no surgery when disease is not present are each favorable (0.80), undergoing unnecessary surgery has half the value (0.40), and that deferring surgery when disease is present has the least value (0.20). The probability threshold would be 0.40, considerably less than the probability that symptoms are attributable to biliary dyskinesia (0.75), so cholecystectomy is advantageous. For a patient with atypical symptoms, no gallstones, and a 35% probability that symptoms are attributable to biliary dyskinesia, cholecystectomy would be disadvantageous.

Djulbegovic et al[36] applied this process to the prophylaxis of deep vein thrombosis (DVT), demonstrating that patients with a DVT risk of 15% or more should receive DVT prophylaxis, and patients with less than 15% risk should not. This approach mandates binary outcome predictions. For models predicting risk along a continuum (ie, 0%–100%), conversion to a dichotomous threshold sacrifices precision, but decision curve analysis obviates conversion to a binary outcome threshold.[37] Decision curve analysis proceeds by solving a decision tree for *pt*, identifying the number of true-positive and false-positive results according to *pt*, calculating the net benefit of the prediction model used to estimate p, and varying *pt* over a clinically relevant range of possible values. Model net benefit is calculated for each new *pt*, producing a decision curve that plots *pt* against a model net benefit for 2 patient populations—one in which all patients have the condition being predicted and one in which no patients have the condition being predicted. The model is beneficial at all *pt* for which the space between the 2 lines has net benefit >0. By avoiding conversion of continuous probability scores to binary variables, this approach has the theoretic advantage of preserving precision.

The tendency to overtreat the insured and undertreat the uninsured in the United States suggests that current practices for incorporating costs in medical decisions are suboptimal.[1] Optimizing value of care, that is, clinical outcomes in the context of financial costs, could address this problem.[38] Decision analyses can accomplish this goal by comparing gains, expressed as quality-adjusted life years (QALYs), with expenditures expressed in monetary values like dollars.[39] Among patients with nonvalvular atrial fibrillation with at least 1 risk factor for stroke, administration of warfarin costs about $8,000 per one QALY saved. For a 65-year old patient with no risk factors, administration of warfarin costs about $370,000 per 1 QALY saved.[40] Robbins et al[41] demonstrate a method for surveying involved parties and incorporating their willingness to bear financial burdens in NNT analyses.

Reinforcement learning

Reinforcement learning is potentially useful in surgical decision-making because it can use an expanded set of complex input data, including text, image, and waveform data tailored to individual patients, to recommend specific actions at sequential decision points. Reinforcement learning is the subfield of artificial intelligence that most closely mimics human learning and decision-making. In this discussion, the agent (an algorithm) learns to map states (patient conditions such as stages of cancer) observed from its environment (data available to the algorithm [eg, data from an electronic health record or a database]) to actions that maximize a

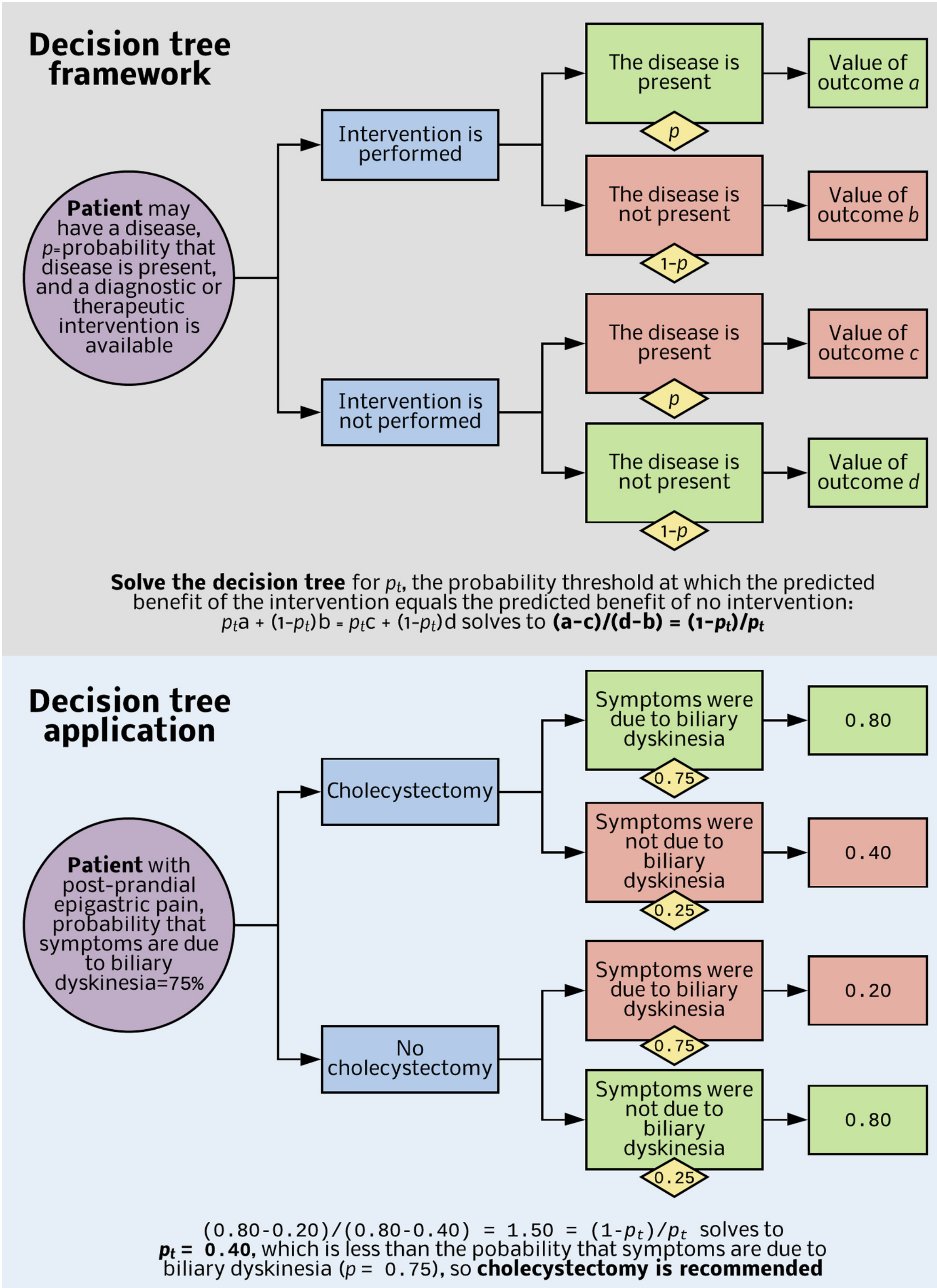

Fig 2. Decision tree framework and clinical application. When it is unclear whether a diagnostic or therapeutic intervention is useful, decision tree analysis identifies a probability threshold (p_t) at which value-adjusted outcomes for intervention and no intervention are equivocal. A prediction model or published literature provides the probability that disease is present. If this value is greater than p_t, then the intervention is useful. Published literature and patient interviews provide relative values for each outcome.

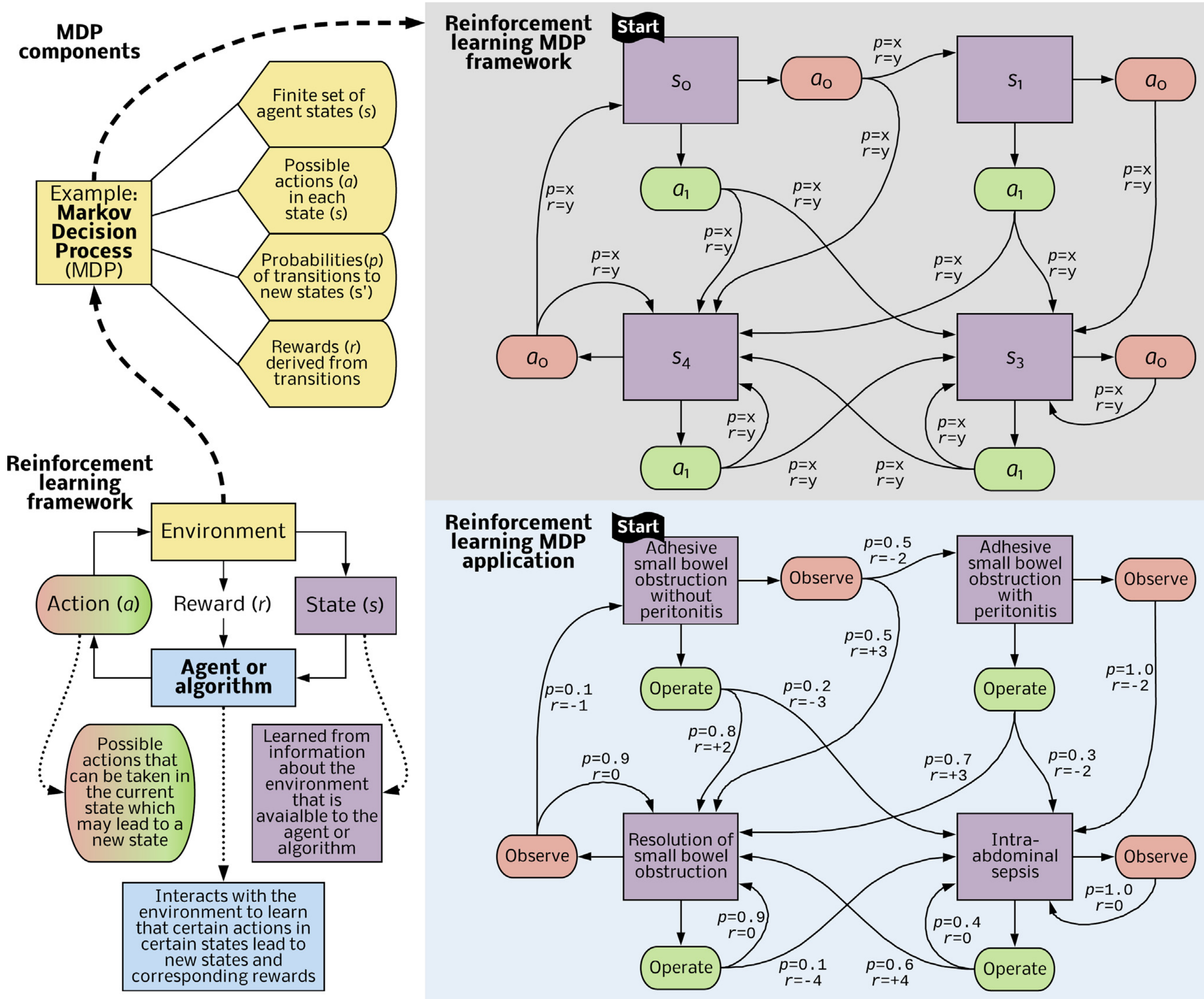

Fig 3. Reinforcement learning framework and clinical application. An algorithm interacts with its environment (consisting of data from electronic health records or data sets) to learn states (representing disease or patient acuity), actions that lead to new states, probabilities of transitioning between states, and associations between state transitions and an ultimate goal, such as survival or discharge to home in good health. The algorithm then identifies actions that are most likely to achieve the ultimate goal. This process can occur within a Markov Decision Process framework and apply to a patient presenting with bowel obstruction, estimating the clinical utility of observation and operative exploration in response to evolving clinical conditions.

reward (clinical outcome). Actions may affect not only the immediate outcomes but also all subsequent states and outcomes.[42] By developing optimal value functions and decision-making policies, reinforcement learning identifies sequences of actions yielding the greatest probability of long-term favorable outcomes as conditions of uncertainty evolve over time. Interactions between a learning algorithm and its environment often occur within a Markov Decision Process containing states, actions, state-transition probabilities, and rewards (Fig 3).

For a patient presenting with adhesive small bowel obstruction without peritonitis, a surgeon may recommend 1 of 2 primary actions: observation or operative exploration. Resolution and discharge home without the need for abdominal exploration, bowel resection, or intra-abdominal sepsis during hospitalization is the goal; however, the goal could be any patient-centered outcome that available data can represent. This “thought experiment” assumes that initial observation yields a 50% chance of transitioning to a state of resolved bowel obstruction, a +3 reward, and a 50% chance that the patient will develop peritonitis, a −2 reward. If instead the patient undergoes early operative exploration, there is an 80% chance of resolution, representing a +2 reward, and a 20% chance of intra-abdominal sepsis attributable to missed enterotomy or surgical site infection, representing a −3 reward. At the next decision step, the patient with intra-abdominal sepsis may be observed, yielding a 100% probability of persistent, intra-abdominal sepsis, or undergo reoperation, yielding a 60% chance of clinical improvement with resolution of obstruction and infection, representing a +4 reward. The algorithm performs a series of such interactions with the environment. The environment sends rewards at each time step, and a value function determines which sequence of actions maximizes the cumulative long-term reward, generating a policy for choosing actions in each state, but also adapting to uncertain conditions that evolve over time. Details regarding “reward” and “value functions” are beyond the scope of this review. We refer interested readers to foundational work on these topics by Sutton and Barto.[42]

Electronic health records

Like other artificial intelligence subfields, most reinforcement learning algorithms require large data sets for training and validation. To achieve the granularity necessary for precise application to individual patients, data sets must be large enough that they contain data from multiple patients that closely mimic the individual patient for whom the decision-support tool is being applied. Many electronic health records (EHR) contain massive quantities of data. Most EHR platforms are adept for billing and ensuring completeness of records, but their interfaces are often cumbersome, and clinically important information lies buried in layers of auto-populated fields. In 1 observational study, medical interns spent 43% of their time during an inpatient rotation using EHRs.[43] Of their time, 13% included direct patient care, down from 25% 2 decades ago.[43,44] One might expect decision-support tools requiring manual data acquisition and entry to be overlooked. Among studies investigating barriers to effective, shared decision-making, time constraint was the most common barrier.[18] In a survey of trainees at academic hospitals, only 26% of all respondents regularly used a risk calculator or other risk-assessment tool.[11] Respondents identified lack of integration with clinical workflow as a major barrier to clinical adoption.

Theoretically, reinforcement learning can capitalize on large data sets in EHRs and obviate manual data entry.[45] It is also possible to expand the input of data for the model to learn from images on radiographs and video monitors and by natural language processing from notes written by clinicians through integration with deep learning, which is adept at parsing large data sets and various types of complex input data. For example, information from computed tomography; cardiac telemetry waveforms; and written descriptions of diseases, operations, and postoperative complications could be processed and represented by deep learning models, and then used as input data for models of reinforcement learning. This approach, “deep reinforcement learning,” has the potential to make the best possible recommendations by incorporating more data requiring no manual input from more sources.

Deep reinforcement learning

For health care applications to be useful, reinforcement learning platforms must efficiently process large volumes of complex data. As the number of variables representing states increases linearly, the combinations and mixtures of data that could represent unique states increase exponentially, computational requirements increase exponentially, and it becomes impossible for naïve or shallow models to perform an exhaustive search for the best possible action in a given state.[46,47] To address this challenge, deep learning and reinforcement learning may be combined, that is, reinforcement learning with parametric function approximation by deep neural networks that efficiently extract key features and patterns from complex environments.[48] When deep learning models are provided with the same data of vital signs and laboratory evaluations used to calculate a traditional illness severity score (for instance the sequential organ failure assessment [SOFA] score often used in an ICU setting), the deep model makes more accurate predictions of mortality.[49] Deep models have performed well in predicting protein structure from raw protein sequences and the impact of human gene mutations.[50,51] Deep models are also adept at tasks that involve computer vision which use pixels as input data to classify images. This technology can apply to radiographs and data from video monitors, expanding the set of input data available to represent environments in reinforcement models.[52–55]

The gaming industry has applied deep reinforcement learning with impressive results. Go is a complex game. There are 32,490 possible first moves, and the number of possible board configurations and available moves increases rapidly as the game progresses. Therefore, an exhaustive search for the optimal move in a certain board configuration with reinforcement learning alone is not feasible. By combining deep and reinforcement learning, an AlphaGo program defeated the European Go champion 5 games to 0.[46] A subsequent version, AlphaGo Zero, was trained purely with deep reinforcement learning using self-play, with no supervised human data and domain knowledge.[56] AlphaGo Zero defeated the earlier version 100 games to 0.

Health care applications

Evidence from retrospective studies suggests that reinforcement learning can apply to clinical decision support. Sepsis is a common, morbid condition for which management strategies are evolving. Within the last decade, evidence-based guidelines have recommended intravenous fluid resuscitation targeting the establishment and maintenance of a central venous pressure of 8 to 12 mm Hg, among other hemodynamic goals. Adherence to this recommendation was associated with administration of nearly 17 L of intravenous fluid within the first 3 days of treatment.[57,58] Unfortunately, sepsis-associated vasoplegia, capillary leak, and decreased ventricular compliance portend poor fluid responsiveness.[59] Less than half of all septic patients with hypotension are fluid responsive, similar to other populations of critically ill patients.[60,61] Excessive administration of intravenous fluid can be harmful. Even among healthy volunteers, only 15% of a fluid bolus remains intravascular 3 h after administration.[62] Fluid boluses, increased central venous pressure, and positive fluid balance have been associated with increased mortality among sepsis patients.[63,64] Methods to ensure optimal balance between intravenous fluid resuscitation and vasopressor administration for patients with sepsis and septic shock remain highly controversial.

Komorowski et al[45] created the AI (Artificial Intelligence) Clinician, a clinical-decision support model capable of recommending the appropriate volume of intravenous fluid and the appropriate doses of vasopressor for septic patients. The model uses a Markov decision process framework in which 90-day survival is the ultimate goal. The model was trained with data from the Medical Information Mart for Intensive Care-III from 61,532 ICU admissions and validated on the Philips eRI data (Koninklijke Philips NV, Amsterdam, The Netherlands) from more than 3.3 million ICU admissions. A total of 48 variables, including vital signs, laboratory values, and comorbidities, were tracked along 4-hour increments during 72 hour and clustered into 750 distinct states. The model “learned” that certain combinations of intravenous fluids and vasopressors were associated with transitions between states, and that certain state transitions were associated with the greatest probability of survival. AI Clinician tended to recommend lesser intravenous fluid and greater doses of vasopressors than clinicians. Mortality was least when actions taken by clinicians matched recommendations from AI Clinician.

When epileptic seizures do not respond to medications, electrical stimulation of the brain and vagus nerve with implantable devices may be a viable alternative treatment. The optimal approach would provide enough neurostimulation to decrease or eliminate seizure activity and minimize cell damage attributable to excessive neurostimulation. The optimal approach is difficult to achieve, attributable in part to difficulties in accurately representing this paradigm with traditional statistical methods and regression modeling. Pineau et al[65] developed a reinforcement learning model to perform this task. The model trained on experimental recordings of in vitro electroencephalogram field potentials that were hand-labeled as normal or seizure activity used to define

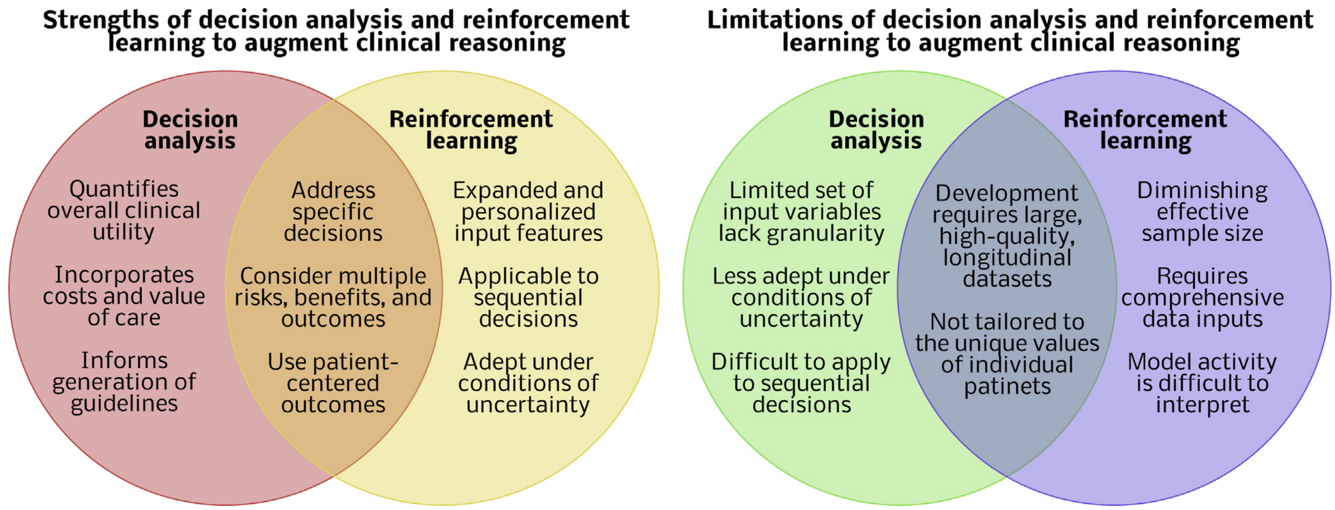

Fig 4. Comparison of decision analysis and reinforcement learning for augmenting clinical reasoning. The unique strengths and limitations of decision analysis and reinforcement learning suggest complementary roles in augmenting clinical reasoning.

various states. Actions included no stimulation or stimulation at three different fixed frequencies. Whereas Komorowski et al[45] targeted a single binary outcome (ie, survival), the Pineau et al[65] study targeted two outcomes (ie, seizure activity and neurostimulation), penalizing both. Minimization of seizure activity was assigned a greater value than the minimization of stimulation, which reflects the clinical observation that seizures are worse than neurostimulation from implantable devices. When applied to experimental data, the model produced decreases in seizure activity comparable to traditional periodic stimulation at fixed frequencies, but with less neurostimulation, thereby achieving the ultimate goal.

Strengths and limitations of decision analysis and reinforcement learning

Decision analysis and reinforcement learning have unique and shared strengths and limitations (Fig 4). These similarities and differences suggest complementary roles in augmenting clinical reasoning across populations and for individual patients.

Strengths

In summary, decision analysis methods quantify overall clinical utility by weighing risks and benefits by patient values and incorporating costs to quantify the "value" of care, facilitating optimal use of resources across health care systems. Probability thresholds inform guidelines and recommendations for diagnostic and therapeutic interventions across populations.[66] Reinforcement learning can use an expanded set of complex input data, including text, image, and waveform data, tailored to individual patients to recommend actions at sequential decision points with uncertain conditions. Both reinforcement learning and decision analysis can make specific recommendations for discrete choices, incorporating multiple risks, benefits, and alternatives of possible interventions and the likelihood that they will lead to patient-centered outcomes of interest.

Limitations

Decision trees and curves typically use few input variables, limiting their ability to represent the unique physiology of individual patients. Like all models, they are less effective when the index patient differs from the cohort used for the development of the model.[15] The same phenomenon limits evidence-based guidelines.[67,68] In addition, decision analyses adapt poorly to conditions of uncertainty, because these decision analyses require that inputs be known or imputed. Finally, the use of simple decision tree and curve analysis is difficult to apply to sequential decision-making, which is often necessary for health care applications.[14]

Reinforcement learning can perform sequential decision-making tasks, but with each additional decision, a smaller proportion of the original sample remains, decreasing the effective sample size.[69] For many surgical diseases, there are no large databases containing all information necessary to solve certain problems with reinforcement learning. Sharing EHR data among institutions could solve this problem but ensuring the interoperability and security of multi-institutional EHR data is difficult both logistically and technically. In addition, when comparing a reinforcement learning policy with clinician decisions, model input data should include all data that truly can influence clinician decision-making.[45] For example, a model recommending operative versus nonoperative management of acute appendicitis should incorporate evidence present on computed tomography of a pericecal phlegmon, suggesting a greater likelihood of the need to perform a greater-risk operation like an ileocecectomy or right hemicolectomy, a greater likelihood that surgeons will recommend nonoperative management, and worse outcomes regardless of management strategies. A model that ignores any pericecal phlegmon could make erroneous associations between nonoperative management and worse outcomes for these patients. Similarly, a model that ignores appendicoliths, which suggest greater likelihood of failing nonoperative management, may underestimate the benefits of early appendectomy for these patients. In these clinical scenarios, the findings on physical examination can make important contributions to surgical decision-making but cannot be included in predictive analytic models with current technologies. Finally, even when all relevant input data are incorporated, it can be difficult to understand how a model reached its recommendation. To mitigate this challenge, methods to improve the transparency and interpretability of the models are available, such as methods that identify model inputs that made important contributions in determining model outputs.[49,70]

Patients and surgeons will want to know how confident the models really are that predictions made by the model will match true, observed outcomes. This need for confidence in the model and suggestions of treatment are important because confidence levels of the machine-learning model can be approximated

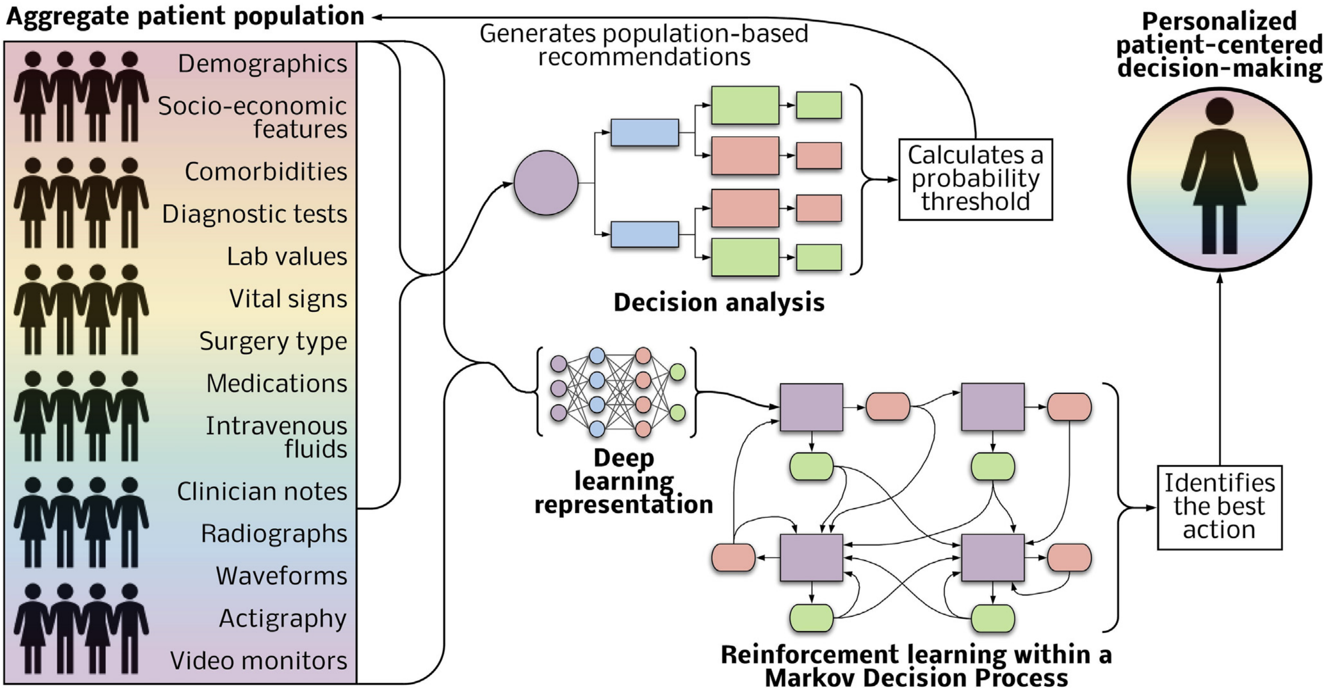

Fig 5. Decision analysis and deep reinforcement learning have complementary roles in augmenting population-based and personalized decision-making. Input variables from patient assessments and the data from the electronic health record feed decision analysis tools that calculate probability thresholds to inform population-based recommendations. Reinforcement learning models combined with deep learning representation of an expanded set of input data can identify actions yielding the greatest probability of a patient-centered outcome.

mathematically to (0.1), with greater values suggesting greater confidence that the model output is accurate, but this method may also overestimate model certainty.[71] Alternatively, predicted probabilities can be calibrated with reliability curves, producing confidence scores.[72] Calibration compares model outputs with a gold standard and answers the question, “Do x of 100 patients with predicted risk x% have the outcome?” This may be depicted graphically or described with the Brier score (calculated as the difference between predicted probability and the actual outcome, raised to the second power), observed-to-expected ratios, or the p value of the Hosmer-Lemeshow goodness-of-fit χ^2 statistic. In predictive analytic terms, calibration compares model predictions with actual outcomes For example, if a perfectly calibrated model predicts a 5% chance of postoperative delirium for 100 different patients, delirium will actually occur in 5 of those patients. Whereas stable discrimination or accuracy depends on consistent effects of the measured covariates on outputs, stable calibration requires that unmeasured covariates make minimal impact on the outcome of interest.[73] Therefore, the performance of the model should be described with both discrimination and calibration. Calibration has a clinically important impact on medical decision-making.[74] Unfortunately, calibration is often omitted in development and validation of models of machine-learning.[72]

Both decision analysis and reinforcement learning require large, high-quality data sets for development and validation. For a patient with early stage breast cancer, the choice to pursue breast-conserving therapy with partial mastectomy and adjuvant radiotherapy limits future treatment options involving additional radiotherapy, which may affect a small subgroup of patients who will develop conditions for which additional radiotherapy is potentially beneficial. Provided with enough granular, longitudinal data, a model could make predictions that consider these subtleties, but such data are often unavailable. EHRs are notorious for noisy data which compromise the performance of traditional and machine-learning models alike.[75] Even when large, longitudinal, high-quality data are available, contemporary approaches to decision analysis and reinforcement learning cannot tailor recommendations to the unique values of individual patients. There may come a time when the availability of massive volumes of data and computational power allows for the efficient training of reinforcement learning models designed to achieve a specific goal that is determined through a shared decision-making process among patients, caregivers, and clinicians. Until then, however, attentive clinicians that understand and interpret clinical context must perform this task. Currently, there is no evidence demonstrating that reinforcement learning can improve surgical decision-making for individual patients or that reinforcement learning is superior to other decision-support methods. Therefore, its potential advantages, though promising, remain theoretic.

Discussion

The unique strengths and limitations of decision analysis and reinforcement learning suggest complementary roles in augmenting clinical reasoning. Decision analysis is well-suited for generating population-based recommendations that optimize clinical utility and value of care. Reinforcement learning is also potentially ideal for individual, patient-centered, sequential decision-making (Fig 5). To produce general recommendations, data from aggregate patient populations regarding the risks and benefits of elective repair of a symptomatic ventral hernia may be considered within the context of financial costs and patient-centered outcomes like long-term functional status and quality of life. In isolation, this may not ensure optimal decision-making for individual patients. Approximately half of all evidence-based practices are provided to patients in the United States.[76] Personalized approaches may succeed where dissemination of clinical practice guidelines has failed. Theoretically, for a patient presenting with a symptomatic ventral hernia, deep reinforcement learning can incorporate an expanded set of input data to determine whether elective repair or expectant management is more likely to yield optimal long-term functional status and quality of life in that specific patient, with sequential recommendations that evolve with changes in clinical conditions over time.

Clinical adoption of reinforcement learning would inevitably lead to disagreements between clinicians and recommendations by the model. There could be substantial legal consequences in assigning liability for adverse events. The nature of the decision also has important implications. Humans and computers both make errors, but patients and their caregivers may have markedly different perceptions regarding human and computer errors regarding sensitive decisions such as situations in which determining futility of care can lead to suggestions of withdrawal of life-sustaining treatments. Finally, a model trained with data from a homogeneous patient population may not represent accurately a separate population or individual patient. For instance, Awad et al[77] reported substantial cross-cultural variation in preferences for moral dilemmas facing self-driving cars. Similar variations likely exist among surgical patients and their caregivers.

In conclusion, surgical patients incur preventable harm from cognitive and judgment errors made under time constraints and uncertainty regarding a patient's diagnosis and predicted response to treatment. Clinicians often ignore or are ignorant of the availability of decision-support tools, which require time-consuming manual entry of appropriate data and lack precision for representing individual patient pathophysiology and clinical context. To address these challenges, decision analysis methods can generate population-based recommendations that jointly consider risks, benefits, costs, and patient values. Reinforcement learning offer the possibility of using large sets of complex patient-specific input data (when available) to identify actions yielding the greatest probability of achieving a goal following a sequence of events as uncertain conditions evolve, offering theoretic advantages for personalized, patient-centered decision-making. The unique potential strengths and limitations of decision analysis and reinforcement learning suggest complementary roles in achieving the ultimate goal of delivering high-value surgical care through sound judgment and optimal decision-making.

Funding/Support

T.J.L. was supported by a postgraduate training grant (T32 GM-008721) in burns, trauma, and perioperative injury from the National Institute of General Medical Sciences (NIGMS). P.T.J. was supported by R01GM114290 from the NIGMS. P.R. was supported by CAREER award, NSF-IIS 1750192, from the National Science Foundation Division of Information and Intelligent Systems, and by the National Insitutes of Health National Institute of Biomedical Imaging and Bioengineering R21EB027344-01. A.B. and P.R. were supported by R01 GM110240. A.B. and P.A.E. were supported by P50 GM-111152 from the NIGMS.

Conflict of interest/Disclosure

The authors have no relevant personal or financial conflicts of interest to disclose.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.1016/j.surg.2020.04.049.

References

1. Hadley J, Steinberg EP, Feder J. Comparison of uninsured and privately insured hospital patients. Condition on admission, resource use, and outcome. *JAMA*. 1991;265:374–379.
2. Goldenson RM. *The encyclopedia of human behavior; psychology, psychiatry, and mental health*. Garden City, NY: Doubleday; 1970.
3. Dijksterhuis A, Bos MW, Nordgren LF, van Baaren RB. On making the right choice: the deliberation-without-attention effect. *Science*. 2006;311:1005–1007.
4. Bekker HL. Making choices without deliberating. *Science*. 2006;312:1472: author reply 1472.
5. Wolf FM, Gruppen LD, Billi JE. Differential diagnosis and the competing-hypotheses heuristic. A practical approach to judgment under uncertainty and Bayesian probability. *JAMA*. 1985;253:2858–2862.
6. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med*. 2005;165:1493–1499.
7. Kirch W, Schafii C. Misdiagnosis at a university hospital in 4 medical eras. *Medicine (Baltimore)*. 1996;75:29–40.
8. Sonderegger-Iseli K, Burger S, Muntwyler J, Salomon F. Diagnostic errors in three medical eras: a necropsy study. *Lancet*. 2000;355:2027–2031.
9. Healey MA, Shackford SR, Osler TM, Rogers FB, Burns E. Complications in surgical patients. *Arch Surg*. 2002;137:611–617; discussion 617-618.
10. Shanafelt TD, Balch CM, Bechamps G, et al. Burnout and medical errors among American surgeons. *Ann Surg*. 2010;251:995–1000.
11. Leeds IL, Rosenblum AJ, Wise PE, et al. Eye of the beholder: risk calculators and barriers to adoption in surgical trainees. *Surgery*. 2018;164:1117–1123.
12. Brotman DJ, Walker E, Lauer MS, O'Brien RG. In search of fewer independent risk factors. *Arch Intern Med*. 2005;165:138–145.
13. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical risk is not linear: Derivation and validation of a novel, user-friendly, and machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) calculator. *Ann Surg*. 2018;268:574–583.
14. Bergquist SL, Brooks GA, Keating NL, Landrum MB, Rose S. Classifying lung cancer severity with ensemble machine learning in health care claims data. *Proc Mach Learn Res*. 2017;68:25–38.
15. Groopman JE. *How doctors think*. Boston: Houghton Mifflin; 2007.
16. Kopecky KE, Urbach D, Schwarze ML. Risk calculators and decision aids are not enough for shared decision making. *JAMA Surg*. 2019;154:3–4.
17. Weinstein JN, Clay K, Morgan TS. Informed patient choice: patient-centered valuing of surgical risks and benefits. *Health Aff (Millwood)*. 2007;26:726–730.
18. Legare F, Ratte S, Gravel K, Graham ID. Barriers and facilitators to implementing shared decision-making in clinical practice: update of a systematic review of health professionals' perceptions. *Patient Educ Couns*. 2008;73:526–535.
19. Bertrand PM, Pereira B, Adda M, et al. Disagreement between clinicians and score in decision-making capacity of critically ill patients. *Crit Care Med*. 2019;47:337–344.
20. de Mik SML, Stubenrouch FE, Balm R, Ubbink DT. Systematic review of shared decision-making in surgery. *Br J Surg*. 2018;105:1721–1730.
21. Wilson A, Ronnekleiv-Kelly SM, Pawlik TM. Regret in surgical decision making: a systematic review of patient and physician perspectives. *World J Surg*. 2017;41:1454–1465.
22. Acuna SA, Chesney TR, Baxter NN. Incorporating patient preferences in non-inferiority trials. *JAMA*. 2019. https://doi.org/10.1001/jama.2019.7059.
23. Giuliano AE, Hunt KK, Ballman KV, et al. Axillary dissection vs no axillary dissection in women with invasive breast cancer and sentinel node metastasis: a randomized clinical trial. *JAMA*. 2011;305:569–575.
24. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA*. 2001;285:2864–2870.
25. McGee TM. Acute appendicitis in pregnancy. *Aust N Z J Obstet Gynaecol*. 1989;29:378–385.
26. Kahneman D, Slovic P, Tversky A. *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press; 1982.
27. Grootendorst P, Feeny D, Furlong W. Health Utilities Index Mark 3: Evidence of construct validity for stroke and arthritis in a population health survey. *Med Care*. 2000;38:290–299.
28. Glasziou PP, Bromwich S, Simes RJ. Quality of life six months after myocardial infarction treated with thrombolytic therapy. AUS-TASK Group. Australian arm of International tPA/SK Mortality Trial. *Med J Aust*. 1994;161:532–536.
29. Solomon NA, Glick HA, Russo CJ, Lee J, Schulman KA. Patient preferences for stroke outcomes. *Stroke*. 1994;25:1721–1725.
30. Man-Son-Hing M, Laupacis A, O'Connor A, et al. Warfarin for atrial fibrillation. The patient's perspective. *Arch Intern Med*. 1996;156:1841–1848.
31. Guyatt GH, Sinclair J, Cook DJ, Glasziou P. Users' guides to the medical literature: XVI. How to use a treatment recommendation. Evidence-Based Medicine Working Group and the Cochrane Applicability Methods Working Group. *JAMA*. 1999;281:1836–1843.
32. Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation. Analysis of pooled data from five randomized controlled trials. *Arch Intern Med*. 1994;154:1449–1457.
33. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med*. 1980;302:1109–1117.

34. Carr JA, Walls J, Bryan LJ, Snider DL. The treatment of gallbladder dyskinesia based upon symptoms: results of a 2-year, prospective, nonrandomized, concurrent cohort study. *Surg Laparosc Endosc Percutan Tech*. 2009;19:222–226.
35. Wybourn CA, Kitsis RM, Baker TA, Degner B, Sarker S, Luchette FA. Laparoscopic cholecystectomy for biliary dyskinesia: which patients have long term benefit? *Surgery*. 2013;154:761–767; discussion 767-768.
36. Djulbegovic B, Hozo I, Lyman GH. Linking evidence-based medicine therapeutic summary measures to clinical decision analysis. *MedGenMed*. 2000;2:E6.
37. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26:565–574.
38. Lee VS, Kawamoto K, Hess R, et al. Implementation of a value-driven outcomes program to identify high variability in clinical costs and outcomes and association with reduced cost and improved quality. *JAMA*. 2016;316:1061–1072.
39. O'Brien BJ, Heyland D, Richardson WS, Levine M, Drummond MF. Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. B. What are the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA*. 1997;277:1802–1806.
40. Gage BF, Cardinalli AB, Albers GW, Owens DK. Cost-effectiveness of warfarin and aspirin for prophylaxis of stroke in patients with nonvalvular atrial fibrillation. *JAMA*. 1995;274:1839–1845.
41. Robbins JM, Tilford JM, Jacobs RF, Wheeler JG, Gillaspy SR, Schutze GE. A number-needed-to-treat analysis of the use of respiratory syncytial virus immune globulin to prevent hospitalization. *Arch Pediatr Adolesc Med*. 1998;152:358–366.
42. Sutton RS, Barto AG. *Reinforcement learning: an introduction*. 2nd ed. Cambridge, MA: The MIT Press; 2018.
43. Chaiyachati KH, Shea JA, Asch DA, et al. Assessment of inpatient time allocation among first-year internal medicine residents using time-motion observations. *JAMA Intern Med*. 2019;179:760–767.
44. Guarisco S, Oddone E, Simel D. Time analysis of a general medicine service: results from a random work sampling study. *J Gen Intern Med*. 1994;9:272–277.
45. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24:1716–1720.
46. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016;529:484–489.
47. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature*. 2015;518:529–533.
48. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science*. 2006;313:504–507.
49. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Sci Rep*. 2019;9:1879.
50. Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet*. 2018;50:1161–1170.
51. Li Y, Kang H, Ye K, Yin S, Li X. FoldingZero: Protein folding from scratch in hydrophobic-polar model. https://arxiv.org/abs/1812.00967. Accessed September 15, 2019.
52. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med*. 2018;15:e1002686.
53. Davoudi A, Malhotra KR, Shickel B, et al. Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning. *Sci Rep*. 2019;9:8020.
54. Hashimoto DA, Rosman G, Witkowski ER, et al. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Ann Surg*. 2019;270:414–421.
55. Topol EJ. A decade of digital medicine innovation. *Sci Transl Med*. 2019;11. https://doi.org/10.1126/scitranslmed.aaw7610.
56. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. *Nature*. 2017;550:354–359.
57. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N Engl J Med*. 2001;345:1368–1377.
58. Jones AE, Shapiro NI, Trzeciak S, et al. Lactate clearance vs central venous oxygen saturation as goals of early sepsis therapy: a randomized clinical trial. *JAMA*. 2010;303:739–746.
59. Marik P, Bellomo R. A rational approach to fluid therapy in sepsis. *Br J Anaesth*. 2016;116:339–349.
60. Lammi MR, Aiello B, Burg GT, et al. Response to fluid boluses in the fluid and catheter treatment trial. *Chest*. 2015;148:919–926.
61. Biais M, Ehrmann S, Mari A, et al. Clinical relevance of pulse pressure variations for predicting fluid responsiveness in mechanically ventilated intensive care unit patients: the grey zone approach. *Crit Care*. 2014;18:587.
62. Chowdhury AH, Cox EF, Francis ST, Lobo DN. A randomized, controlled, double-blind crossover study on the effects of 2-L infusions of 0.9% saline and plasma-lyte(R) 148 on renal blood flow velocity and renal cortical tissue perfusion in healthy volunteers. *Ann Surg*. 2012;256:18–24.
63. Boyd JH, Forbes J, Nakada TA, Walley KR, Russell JA. Fluid resuscitation in septic shock: a positive fluid balance and elevated central venous pressure are associated with increased mortality. *Crit Care Med*. 2011;39:259–265.
64. Maitland K, Kiguli S, Opoka RO, et al. Mortality after fluid bolus in African children with severe infection. *N Engl J Med*. 2011;364:2483–2495.
65. Pineau J, Guez A, Vincent R, Panuccio G, Avoli M. Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach. *Int J Neural Syst*. 2009;19:227–240.
66. Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol*. 2018;74:796–804.
67. Tinetti ME, Bogardus Jr ST, Agostini JV. Potential pitfalls of disease-specific guidelines for patients with multiple conditions. *N Engl J Med*. 2004;351:2870–2874.
68. Boyd CM, Darer J, Boult C, Fried LP, Boult L, Wu AW. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. *JAMA*. 2005;294:716–724.
69. Gottesman O, Johansson F, Komorowski M, et al. Guidelines for reinforcement learning in healthcare. *Nat Med*. 2019;25:16–18.
70. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable deep models for ICU outcome prediction. *AMIA Annu Symp Proc*. 2016;2016:371–380.
71. Gal Y. Uncertainty in deep learning: Dissertation 2016. University of Cambridge Web site. http://mlg.eng.cam.ac.uk/yarin/thesis/thesis.pdf. Accessed July 13, 2019.
72. Guo C, Pleiss G, Sun Y, Weinberger K. On calibration of modern neural networks. https://arxiv.org/pdf/1706.04599.pdf. Accessed July 13, 2019.
73. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol*. 2010;172:971–980.
74. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making*. 2015;35:162–169.
75. Goldstein BA, Pomann GM, Winkelmayer WC, Pencina MJ. A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis. *Stat Med*. 2017;36:2750–2763.
76. McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. *N Engl J Med*. 2003;348:2635–2645.
77. Awad E, Dsouza S, Kim R, et al. The Moral Machine experiment. *Nature*. 2018;563:59–64.

Surgery is abstracted and/or indexed in *Index Medicus, Science Citation Index, Current Contents/Clinical Medicine, Current Contents/Life Sciences,* and MEDLINE.