original reports

Artificial Intelligence Approach for Variant Reporting

abstract

Purpose Next-generation sequencing technologies are actively applied in clinical oncology. Bioinformatics pipeline analysis is an integral part of this process; however, humans cannot yet realize the full potential of the highly complex pipeline output. As a result, the decision to include a variant in the final report during routine clinical sign-out remains challenging.

Methods We used an artificial intelligence approach to capture the collective clinical sign-out experience of six board-certified molecular pathologists to build and validate a decision support tool for variant reporting. We extracted all reviewed and reported variants from our clinical database and tested several machine learning models. We used 10-fold cross-validation for our variant call prediction model, which derives a contiguous prediction score from 0 to 1 (no to yes) for clinical reporting.

Results For each of the 19,594 initial training variants, our pipeline generates approximately 500 features, which results in a matrix of > 9 million data points. From a comparison of naive Bayes, decision trees, random forests, and logistic regression models, we selected models that allow human interpretability of the prediction score. The logistic regression model demonstrated 1% false negativity and 2% false positivity. The final models' Youden indices were 0.87 and 0.77 for screening and confirmatory cutoffs, respectively. Retraining on a new assay and performance assessment in 16,123 independent variants validated our approach (Youden index, 0.93). We also derived individual pathologist-centric models (virtual consensus conference function), and a visual drill-down functionality allows assessment of how underlying features contributed to a particular score or decision branch for clinical implementation.

Conclusion Our decision support tool for variant reporting is a practically relevant artificial intelligence approach to harness the next-generation sequencing bioinformatics pipeline output when the complexity of data interpretation exceeds human capabilities.

Clin Cancer Inform. © 2018 by American Society of Clinical Oncology

Michael G. Zomnir
Lev Lipkin
Maciej Pacula
Enrique Dominguez Meneses
Allison MacLeay
Sekhar Duraisamy
Nishchal Nadhamuni
Saeed H. Al Turki
Zongli Zheng
Miguel Rivera
Valentina Nardi
Dora Dias-Santagata
A. John Iafrate
Long P. Le
Jochen K. Lennerz

Author affiliations and support information (if applicable) appear at the end of this article.

Corresponding author: Jochen K. Lennerz, MD, PhD, Massachusetts General Hospital/Harvard Medical School, Department of Pathology, Center for Integrated Diagnostics, 55 Fruit St, WRN503, Boston, MA 02114; e-mail: jlennerz@partners.org.

INTRODUCTION

Over the past decade, several areas in clinical oncology have been revolutionized by cancer genotyping.[1-3] And next-generation sequencing (NGS) technologies will continue to play an integral part of precision medicine.[3-6] From an analytic perspective, the amount and complexity of raw NGS data require sophisticated pipelines (Fig 1) for read alignment, variant calling, and annotation, each of which may be associated with error and noise; this is also true for the sequencing instrumentation itself. Ultimately, these pipelines procure a text file containing a list of sequence variants, called a variant call format (VCF) file, wherein each variant is listed with a set of features (ie, annotations). These annotations span hundreds of columns and include functional predictions, variant calling metrics, frequencies in public databases, and/or clinical implications. Molecular pathologists/geneticists interpret these VCF files and integrate selected data points when issuing a clinical report (Fig 1).

Knowledge derived from the literature and public databases provides comprehensive disease-variant associations.[7] By using clinical practice data, we can achieve similar disease-variant associations as exemplified by a two-feature representation (disease sites and reported variants by gene; Fig 1). The final condensed matrix does not, however, fully represent the complex reporting process and does not aid the molecular pathologist/geneticist with the day-to-day, case-by-case, and variant-by-variant reporting decision.

Numerous approaches[2,8-10] and guidelines for variant calling,[11-13] annotation,[12,14,15] and interpretation[8,9,16-18] have been proposed. Ultimately, these strategies share the common end goal of

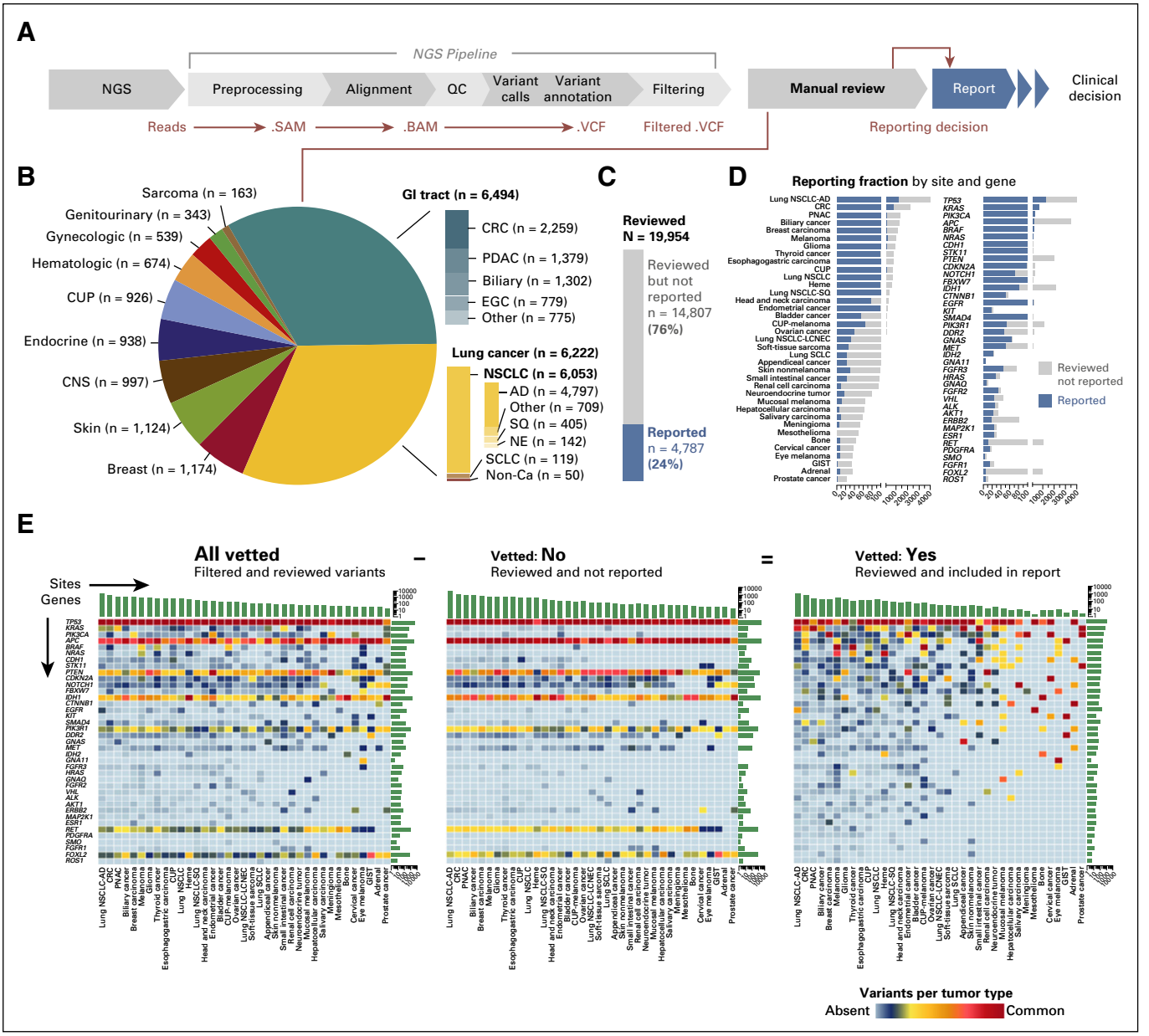

Fig 1. The complexity of variant reporting in clinical practice. (A) The amount and complexity of raw next-generation sequencing (NGS) data requires NGS pipelines for read alignment, variant calling, and variant annotation to provide a (filtered) variant call format (VCF) file for manual review by a pathologist/geneticist. The reporting decision is a complex process that requires experience, involves management of the VCF file and various resources, and ultimately results in a reporting decision. (B) Distribution of tumor types included in the variant training data set (V1). Variants are represented in 37 principal tumor types that combine 383 histologic subtypes. (C) After manual review of 19,594 variants, only 24% (n = 4,787) are reported, and 76% of the review effort is not captured in the final report. (D) The reporting fraction by site (left) and gene (right) shows considerable variation (range, 0% to 100%). (E) The effect of the variant reporting decisions illustrated on a variant frequency matrix; green bars represent the number of variants within each disease site or gene. We used the formula "all" minus "no" equals "yes." Specifically, the filtered pipeline output represents "all" reviewed variants and after subtraction of the variants that received "no" calls (ie, are vetted not to be included in the report), the resulting matrix shows the variant frequencies by gene and site in the final report (ie, "yes" calls). The resulting "yes" matrix is similar to that in recent publications[7]; however, in clinical practice, pathologists/geneticists are confronted with all data ("all" matrix on the left). The portrayed distribution of variants by gene and site represents only two of approximately 500 pipeline features attached to each variant. The full pipeline output and the dimensionality of interrelations exceed the human ability to handle all available data efficiently. AD, adenocarcinoma; BAM, binary alignment map; CRC, colorectal cancer; CUP, carcinoma of unknown primary; EGC, esophagogastric cancer; GIST, GI stromal tumor; Heme, hematologic malignancies; LCNEC, large-cell neuroendocrine carcinoma; NE, neuroendocrine carcinoma; Non-Ca, nonepithelial malignancy; NSCLC, non–small-cell lung cancer; PDAC, pancreatic cancer; QC, quality control; SAM, sequence alignment map; SCLC, small-cell lung cancer; SQ, squamous cell carcinoma.

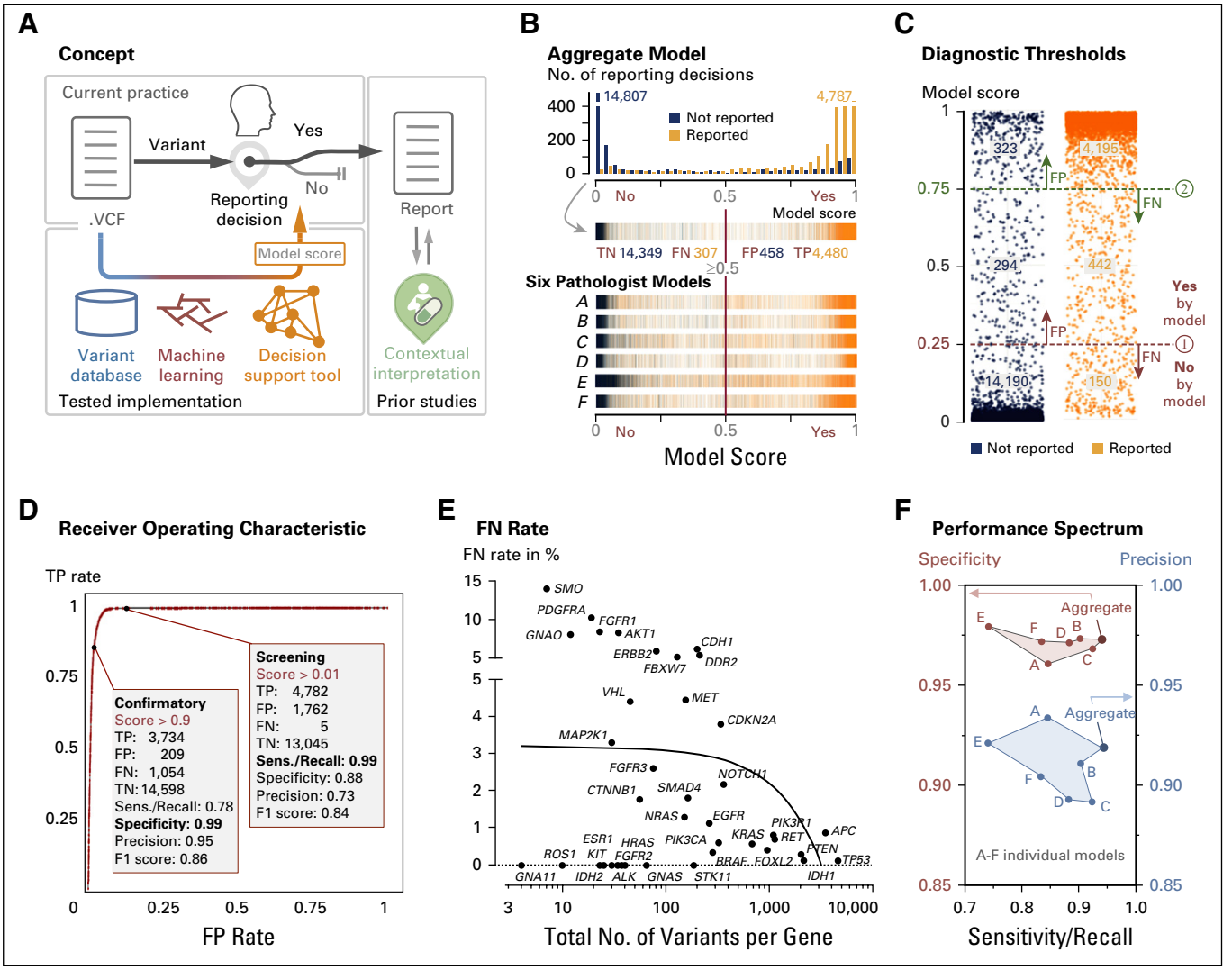

Fig 2. Performance assessment of the artificial intelligence model for variant reporting. (A) Concept of a decision support tool for variant reporting. Current practice (top) is shown with the tested implementation (bottom). The artificial intelligence/machine learning model was built on the basis of prior human reporting decisions. Note that the implemented model provides a reporting decision for each variant on a scale from 0 (no) to 1 (yes) without regard for potential clinical actionability; contextual or clinical consequences (eg, oncology knowledge database) have been excluded intentionally, and we have addressed the topic in prior studies.[16] (B) The number of calls in the aggregate model (by using a naive threshold of 0.5) as well as distribution of no and yes calls per pathologist (A to F). (C) Distribution of 19,954 model scores in the reported and not reported variants. Two call thresholds illustrate two use cases: (1) a more-sensitive 0.25 threshold with fewer false-negative (FN) results (n = 150) and (2) a more-specific 0.75 threshold with fewer false-positive (FP) results (n = 323). (D) Receiver operating characteristic curves with selected performance metrics for model threshold scores of > 0.01 (screening test) and > 0.9 (confirmatory test). (E) The FN rate decreases with increasing prevalence; however, for several genes, the model performance is excellent despite low prevalence (eg, *ALK*). (F) Specificity over sensitivity (red) and precision over recall (blue) for the aggregate and individual models (A to E). The outline of all individual models can be viewed as the model-based performance spectrum of the examined group practice composed of six pathologists (A to E). TN, true negative; TP, true positive; Sens., sensitivity; VCF, variant call format.

the pathologist/geneticist rendering a seemingly basic decision: whether to report a variant. To our knowledge, the binary reporting decision has not been captured or used systematically.[8-10,16,19] In particular, the decision to not report a variant (Figs 1C and 1E; Data Supplement) is equally valuable because the same variant may be found in subsequent cases. In times of escalating costs, building on prior knowledge may save time and effort. Comprehensive pipeline results may help, but the reporting decision remains difficult relative to the context of the called variant (eg, presence of a mutation at low levels, variants at sites of private polymorphisms, unknown functional consequences of variants, splice site variants). To render the reporting decision, a board-certified pathologist/geneticist uses experience with numerous features distributed over many columns to render a reporting decision (called"vetting"; Fig 2A). Some full

pipeline outputs consist of > 500 dimensions (or features) for each variant. In other words, for a clinical NGS laboratory, the integration and processing of the full potential of the entire pipeline output as it relates to the reporting decision exceed human capabilities.[8,19,20]

Artificial intelligence is one approach to mine big data and derive models for decision making.[21-23] Common repetitive tasks in medicine are amenable to modeling by artificial intelligence[21-23]; however, computers are notoriously poor at understanding unstructured data (eg, medical notes).[21,23-25] In contrast, artificial intelligence tools for structured data are available and have already surpassed human performance in many areas.[25-31] Of note, bioinformatics pipelines generate mostly structured, discrete data. Thus, we consider the inability of humans to gain access to the full potential of the pipeline output, coupled with the discrete nature of the data and the final binary reporting decision, as an ideal setting[32] to assess the performance of an artificial intelligence–based decision support system for variant reporting.

On the basis of our experience with implementing clinical genotyping[33,34] that uses NGS technologies,[35] we report on how we aligned big clinical sequencing data with human decisions in our group. As a result, we established an artificial intelligence approach for variant reporting and describe its performance in two independent data sets from routine clinical practice.

METHODS

Design, Regulatory Approval, and Clinical Setting

The project was undertaken as a retrospective analysis of existing data obtained as part of routine cancer care in a clinical molecular genetics laboratory. All patients provided written informed consent for molecular genotyping. Institutional review board approval was obtained (protocol 2014P000940). Details about the project site, case volume, and assays can be found in Zheng et al[35] and the Data Supplement.

NGS Assays and Bioinformatics

For sequence analysis, we used our laboratory-developed, Clinical Laboratory Improvement Amendments–validated NGS bioinformatics pipeline. We used Illumina (San Diego, CA) MiSeq (V1) or NextSeq (V2) instruments. We performed tumor-only sequencing, and the 2 × 151–base paired end sequencing results were aligned to the hg19 human genome reference by using Burrows-Wheeler Aligner MEM.[36] Variant calling for V1 (Data Supplement) was performed with MuTect version 1.1.7 for single-nucleotide variants,[37] and Oncotator (version 1.2.10.0)[38] was used for variant annotation. For V2 (Data Supplement), we used Novoalign (www.novocraft.com) for read alignment and an ensemble variant calling approach, including MuTect1, LoFreq, Genome Analysis Toolkit, and a laboratory-developed hotspot caller, for variant detection. The variant effect predictor tool was used for variant annotation.[39] A detailed standard operating procedure is available upon request; a description of the variant scoring and features used for analysis in V1 and V2 are provided in the Data Supplement.

Machine Learning and Classifier Selection

Python version 2.7.11, Pandas version 0.18.1, NumPy version 1.11.0, and scikit-learn version 0.17.1 libraries (Python Software Foundation, Wilmington, DE) were applied for model building. By following recommendations by Kohavi[40] for real-world data sets, the best method to use for model selection is 10-fold stratified[40-44] cross-validation.[42,43] For selection of the most accurate classifier,[42] we compared area under the curve (AUC) values for naive Bayes, logistic regression, decision trees, random forests (depth of trees, 10, 15, 50, and 100), and support vector machines (scikit-learn library).[45] We refer to artificial intelligence model or model as the selected predictor/classifier for implementation (Data Supplement).

Aggregate and Individual Models

In our clinical practice, each case is signed out by one pathologist (detailed call frequency by gene and disease site across all pathologists are provided in the Data Supplement). For the aggregate model, we trained and tested by using all reporting call decisions as labels (ie, not included as a feature in the model). For the individual models, we filtered the training sets to include only calls made by one pathologist (Fig 2B); the six resulting models (A to F) can be regarded as representative of each individual

Table 1. Overview of the Data Sets

Component	No.	Description
Data set (V1)		
Variants	19,954	Unique variants from clinical practice
Disease sites	383	Unique primary diagnosis from 37 principle disease sites (eg, brain, lung, etc)
Genes	39	Gene names (by following HGNC nomenclature)
Time frame, months	32	November 2013 to June 2016
Data set (V2)		
Variants	16,123	Unique variants from clinical practice
Disease sites	398	Unique primary diagnosis from 37 principle disease sites (eg, brain, lung, etc)
Genes	116	Gene names (by following HGNC nomenclature)
Time frame, months	11	September 2016 to July 2017
Pathologists	6	No. of pathologists

Abbreviation: HGNC, Human Genome (HUGO) Gene Nomenclature Committee.

pathologist's reporting practice. In addition, we assessed transferability of the approach to a new assay (T1/T2 experiments [ie, trained model on variants V1 and tested on V2]) and validated the performance in an independent data set (V2; Data Supplement).

Statistical Analysis

We defined reporting rates of pathologists as the number of yes decisions over the total number of decisions. To account for different use cases, we describe the model performance by using different score cutoffs (naive *v* outliers *v* screening *v* confirmatory; Data Supplement). To assign *P* values to individual features, we used univariable analysis of variance testing (Data Supplement). We assessed the performance of the model by comparing the model-based predictions versus the original clinically reported variant calls and provide precision, recall, sensitivity, specificity, F1 score, and Youden index. Statistical significance was defined as $P < .05$.

RESULTS

Data Extraction Results

An overview of the data set V1 and the validation data set V2 is listed in Table 1. In V1, our pipeline assigns 507 features to each variant (Data Supplement). The reporting pathologist/geneticist could be regarded as a feature; however, this feature was not included because it is not available when confronted with a new variant. We provide however, the number of cases, variants, and yes call rates by pathologist (Table 2). The number of reviewed variants (approximately five per case) and the number of calls per case (fewer than two plus or minus two per case) did not differ significantly (Table 2). In other words, despite some variation in case and/or variant exposure of every pathologist (Table 2; Data Supplement), we considered the call rates to be internally consistent and the aggregate total variant number as representative of our sign-out practice.

Selection of the Machine Learning Classifier

We compared the AUCs for several different predictor methods (Data Supplement). In V1, we selected a logistic regression model (over the equally predictive random forest classifier), and in V2, we selected random forests as the top performing model (Data Supplement). Specifically, although each model derived a single number, we also considered human interpretability; we regard the individual coefficients or decision branches for each feature as an additional benefit for manual review (transparency of the model).

Results From Implementing the Model

The prediction model assigns a score from 0 to 1 that represents a continuous scale from the ground truth reporting calls (no to yes, respectively). Review of the frequency distribution

Table 2. Number of Cases, Variants, and Calls by Pathologist in V1

Pathologist	Cases	Variants	Variants per Case ± SEM	"Yes" Calls (%)	"Yes" Calls per Case ± SEM	*P**
A	723	4,015	5.55 ± 0.083	959 (23.9)	1.91 ± 0.051	.45
B	1,118	6,170	5.52 ± 0.064	1,502 (24.3)	1.86 ± 0.041	.73
C	314	1,765	5.62 ± 0.120	419 (23.7)	1.83 ± 0.070	.38
D	830	4,564	5.50 ± 0.075	1,152 (25.2)	1.90 ± 0.041	.33
E	92	487	5.29 ± 0.200	122 (25.1)	1.77 ± 0.120	.23
F	453	2,593	5.72 ± 0.099	633 (24.4)	1.97 ± 0.069	.23
Total	3,530	19,594	5.55 ± 0.036	4,787 (24.4)	1.89 ± 0.023	NA

Abbreviations: NA, not applicable; SEM, standard error of the mean.

**P* values derived from Fisher's exact tests that compared each pathologist's average yes call rate against all others (eg, A *v* non-A).

of calls for all 19,594 variants by model score showed that the model assigned the majority of scores near 0 and 1 (Fig 2B, top). To allow comparison of how each pathologist contributes to the aggregate model, we plotted individual model scores (Fig 2B; Table 3) and assessed several call thresholds (Figs 2B and C; Data Supplement). When plotting the true-positive rate over the false-positive rate for all existing model score cutoffs as a receiver operating characteristic curve (Fig 2D), the AUC can be interpreted as the probability that a randomly selected positive variant in the test data receives a higher score from the model than a randomly selected negative variant. The resulting receiver operating characteristic curve shows an AUC of 0.990, which indicates an almost perfect prediction model (Table 3, aggregate model V1), and we provide the specific test characteristics and case numbers of a screening cutoff at 0.01 and a confirmatory cutoff at 0.9 (Fig 2D).

Clinical Implementation

We implemented the model for screening (cutoff > 0.01) and observed that the false-negative rate decreases with the overall variant prevalence by gene (Fig 2E), which is related to our call practice (Fig 2B; Data Supplement). For example, *SMO* variants in our setting are overall rare (n = 7)

Table 3. Performance of the Aggregate and Individualized Models

		Performance Measures					
Model	TC	AUC	Precision	Recall or Sensitivity	Specificity	Youden Index	F1 Score*
Aggregate (V1)	19,594	0.989	0.907	0.936	0.969	0.915	0.921
Pathologist							
A	4,015	0.986	0.934	0.845	0.961	0.806	0.887
B	6,170	0.988	0.911	0.904	0.973	0.877	0.907
C	1,765	0.985	0.892	0.824	0.969	0.793	0.857
D	4,564	0.988	0.893	0.883	0.971	0.854	0.888
E	487	0.982	0.921	0.741	0.979	0.720	0.821
F	2,593	0.984	0.905	0.835	0.972	0.807	0.868
Validation set (V2)	16,123	0.987	0.938	0.953	0.974	0.927	0.945
Transferability							
T1 not retrained	568	0.768	0.903	0.567	0.755	0.322	0.697
Δ |T2 – T1|	NA	0.212	0.022	0.356	0.204	0.560	0.204
T2 retrained	568	0.980	0.881	0.923	0.959	0.882	0.901

NOTE. The table provides the performance of the model-based predictions when compared with the originally clinically reported calls (labels). V1 and V2 refer to distinct variant sets (Table 1). V1 is shown in Figs 1 and 2, whereas V2 refers to an entirely separate validation set of variants (September 2016 to July 2017). To quantify the importance of retraining, we performed two transferability experiments: T1 refers to an intentionally wrong application of a not retrained model compared with T2, which is an appropriately retrained model (Data Supplement). Abbreviations: AUC, area under the curve; TC, training call.

*Measure of test accuracy (harmonic mean of precision and recall)

yet frequently reported (n = 5; Data Supplement). As a consequence, the one false-negative call by the model results in a relatively high false-negative rate (15%). However, despite these limitations, the overall error and false-negative rate is only approximately 3%. We used the same approach to derive individual models for each pathologist (Table 3; Figs 2D and F), which illustrates three things. First, the aggregate model outperforms most individual models. Second, the performance among all individual models differs only by 0.8% (AUCs, 0.982 to 0.990). Third, our approach to derive a model for every pathologist suggests that there are only small differences in sign-out practice for variant reporting that are reflected in the six models (akin to a consensus difference). Figure 2F shows the model-based landscape of our variant reporting practice.

Validation Experiments

The launch of a new assay in July 2016 provided an opportunity to transfer our artificial intelligence model from the V1 assay to the new V2 assay, and in doing so, to validate our approach. We used V2 as an independent data set that contained 16,123 consecutive variants from clinical practice (V2; Table 1). The performance measures (Table 3; Data Supplement) and the Youden index of 0.927 validate the machine learning model as robust in capturing our sign-out experience with respect to reporting decisions.

Transferability Experiments

The launch of a new assay also triggered the following question: What would happen to the model performance if we intentionally transferred the wrong model? We consider this transferability question to be highly relevant because any new assay (or laboratory test) will have slightly different frequencies of cancer types, genes, and so forth. As an analogy, when a pathologist/geneticist starts to work with a new assay, he or she would need to retrain and gain familiarity with the assay to confidently sign out cases. We designed transferability experiments to assess the importance of retraining and the risk imposed by wrongful artificial intelligence model transfer (Data Supplement). The lower performance (up to –35%) is no surprise and related to substantially different panels; however, some performance measures are preserved (precision, approximately 0.9; Table 3). These findings indicate that some variants are shared between the assays and that transfer of artificial intelligence models requires careful multiparametric performance evaluation. The validation experiments (V1 *v* V2) confirmed transferability of the machine learning approach, whereas the transferability experiments (T1/T2) assessed the transferability of the model. These experiments emphasize that pathologists' calls are necessary to retrain a model and that the approach can be accurately transferred to capture reporting decisions on an entirely different assay.

Implementation in Clinical Practice

To illustrate how we implemented the models in clinical practice, we provide screenshots of two variant review popup tools. First, a consensus tool (Fig 3A) allows for the review of the reporting prediction scores to augment variant reporting decisions and instantly access the diversity of reporting suggestions along with a drill-down functionality that allows review of the features that contributed to a particular score. These functionalities capture two key aspects of a traditional consensus conference: the decision and the reasoning. Second, we implemented a tree map visualization of the decision branches and the underlying reasoning (Fig 3B). The added level of transparency was a core design component that empowers the reviewing pathologist/geneticist to understand the underlying reasoning; in other words, the model provides the user with a justification for a given decision.

DISCUSSION

We present an artificial intelligence approach for genetic variant reporting in cancer. The core design idea is to use artificial intelligence to fully exploit the entire bioinformatics pipeline output for modeling the reporting decision. We combined the structured variant annotations with the collective, multiyear reporting experience of six pathologists in a clinical molecular diagnostic laboratory. By using > 19,000 variants, we derived substantial performance metrics with AUC > 99%, and we validated the approach in > 16,000 independent variants in another assay with similarly high

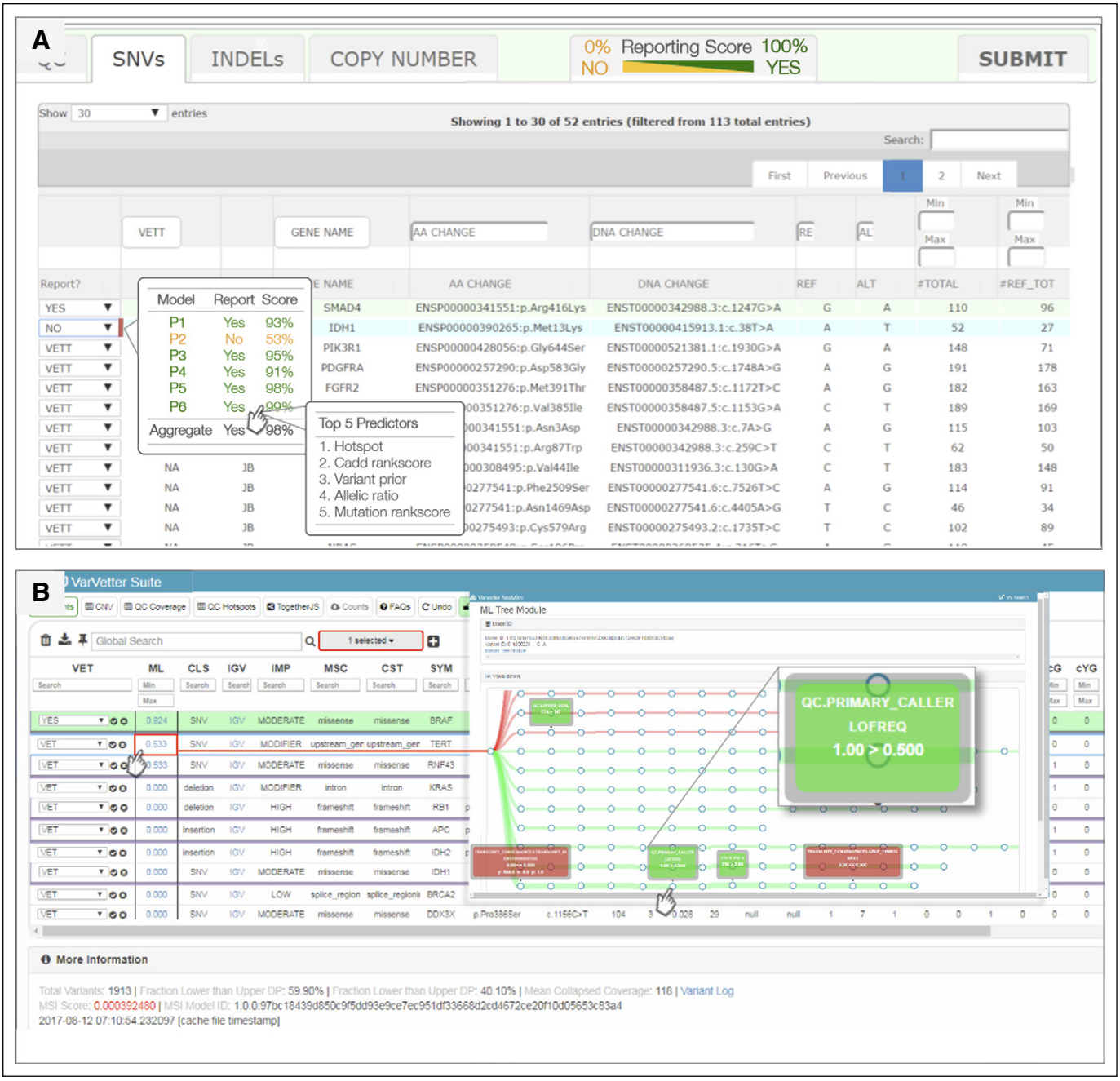

Fig 3. Model decision exploration in clinical practice. (A) Screenshot shows our variant review and graphic user interface used to select variants for inclusion in the report (background; old assay, V1). The inset shows individual pathologists' model scores (P1 to P6) and the aggregate. When hovering over one model, the drill-down option shows the top five predictors derived from the logistic regression pathologist's model that contributed to the report recommendation (report). (B) Screenshot shows our variant review and graphic user interface used to select variants for inclusion in the report (background; new assay, V2). The machine learning (ML) score links out to the ML tree module, which allows for exploration of 15 random forest decision branches. Each branch contains the order of contributing features and findings that resulted in the decision (green argues for reporting, red against). Each circle represents one feature, and the drill-down option (inset) shows the feature (eg, a quality control [QC] metric of a caller), the finding in this variant (eg, 1), and the cutoff used by the model (here > 0.5). The added level of transparency that allows review of the features that underlie a model-derived decision is an important design component of our implementation in clinical practice, and we propose the term next-generation decision support. CADD, combined annotation dependent depletion; LOFREQ, low frequency; SNV, single-nucleotide variant.

performance metrics. We also share two clinically relevant visualization tools (a consensus and a tree map visualization module) that allow human exploration of the underlying (supporting) reasons by the artificial intelligence model. This approach is a practical example of how to apply artificial intelligence meaningfully when the complexity of data exceeds human capabilities.

Artificial intelligence and machine learning as tools have been established,[40,42-45] and in our model design,[32] we aimed for two things: to generate a single score output as a contiguous score from 0 to 1 as a simple way to recognize the decision and to give the reviewing pathologist/geneticist the opportunity not only to review the decision but also to explore the underlying reasoning by the model (Fig 3). We offer this additional layer of transparency as a drill-down function and propose the term next-generation decision support tool. Given that the model takes the full set of variant calling pipeline output into consideration, we foresee review of these features that underlie a reporting decision as valuable (eg, machine-based education of the reviewing pathologist). Unfortunately, when machines or programs mimic components of the cognitive functions that humans attribute to learning, the disparity between intended use and public perception generates unnecessary confusion. For example, in science fiction, the term artificial intelligence typically is applied to excite ethical and emotional reactions. Thus, we want to re-emphasize the critical importance of thoughtful design and that we limited our approach to the binary reporting decision. Thus, the intended use of our model is not to substitute human decision making. Comparison of human versus model decisions enables the pathologist to decide whether the variant should be re-reviewed by an independent pathologist or presented at a consensus conference. Thus, the tool may increase efficiency and reinforce human interaction. In other words, we built, implemented, and validated a model that continues to rely on unbiased pathologist calls; this small integral design part literally means: no pathologist, no model.

Future adoptions may extend to incorporate other data types (eg, histopathologic image features),[46] other artificial intelligence approaches (ie, neural networks, deep learning),[47,48] or the transfer to other laboratories or assays (as demonstrated in our transferability experiments). Despite the overall promising nature of our findings, numerous limitations apply. The model currently does not distinguish somatic from germline variants for prediction, and we did not include other forms of variants (ie, indels, copy number variation). With regard to implementing our approach in other clinical laboratories, limitations exist in terms of different frequencies of diseases or insufficient clinical volumes to drive model building. Our model is restricted to the collective experience of our group (eg, we do not sequence matched normals). Confounders and biases may exist in our model with regard to such nuances as the ethnicity or ancestry of our testing population compared with the reference genome used for variant calling. Furthermore, the launch of a new assay shows evolution of performance over time, and our transferability experiments demonstrate that this also applies to an artificial intelligence model because the model is, by definition, assay, instrument, and pipeline specific. Although these limitations can be perceived as disadvantageous, they are not specific to our own practice and are well acknowledged in the artificial intelligence and machine learning community.[20,23,25,49] Despite these limitations, we have shown that although models may be assay and setting specific (V1 *v* V2; Table 3), the application of artificial intelligence for decision support is indeed a practically feasible and clinically relevant strategy.

To the best of our knowledge, our artificial intelligence model as a decision support tool for variant reporting is new. Given the performance metrics of the model, appropriate model cutoffs (< 0.01) significantly reduce the number of variants to be reviewed (Fig 2D). Thus, we view the deployment of an artificial intelligence model in a clinical environment as a tool to filter out nonreportable variants and to enable exploration of the underlying reasoning as highly relevant because it makes pipeline result vetting reasonable timewise and scalable in terms of content and for increased numbers of clinical samples.

The implicit understanding that big data efforts will ultimately improve decision making in health care entails optimization of the care coordination process. We consider the complexity of NGS bioinformatics pipeline outputs as an opportunity to apply artificial intelligence because the complexity exceeds human capabilities. In times of increasing health care costs, tools for increased efficacy are in demand, and we hope that by sharing our approach to leverage the power of artificial intelligence and apply it to genetic variant reporting in cancer, other groups will

be motivated to explore this emerging field and improve patient care.

DOI: https://doi.org/10.1200/CCI.16.00079
Published online on ascopubs.org/journal/cci on March 22, 2018.

AUTHOR CONTRIBUTIONS

Conception and design: Michael G. Zomnir, Maciej Pacula, Enrique Dominguez Meneses, Nishchal Nadhamuni, A. John Iafrate, Long P. Le, Jochen K. Lennerz

Collection and assembly of data: Michael G. Zomnir, Lev Lipkin, Maciej Pacula, Enrique Dominguez Meneses, Allison MacLeay, Sekhar Duraisamy, Saeed H. Al Turki, Miguel Rivera, A. John Iafrate, Long P. Le, Jochen K. Lennerz

Data analysis and interpretation: Jochen K. Lennerz, Lev Lipkin, Sekhar Duraisamy, Zongli Zheng, Valentina Nardi, Dora Dias-Santagata, A. John Iafrate, Long P. Le

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Michael G. Zomnir
No relationship to disclose

Lev Lipkin
Stock and Other Ownership Interests: TEVA Pharmaceuticals Industries, Pfizer, Novartis

Maciej Pacula
Patents, Royalties, Other Intellectual Property: Ute Geigenmuller, Doris Damian, Maciej Pacula, Mark A. DePristo. Methods and Systems for Determining Autism Spectrum Disorder Risk (US patent 9,176,113), granted November 3, 2015 (Inst)

Enrique Dominguez Meneses
No relationship to disclose

Allison MacLeay
Travel, Accommodations, Expenses: InterSystems, Athenahealth (I)

Sekhar Duraisamy
No relationship to disclose

Nishchal Nadhamuni
No relationship to disclose

Affiliation

All authors: Massachusetts General Hospital, Boston, MA.

Saeed H. Al Turki
Honoraria: Foundation Medicine
Consulting or Advisory Role: Pfizer
Travel, Accommodations, Expenses: AstraZeneca, Foundation Medicine

Zongli Zheng
Stock and Other Ownership Interests: Archer Biosciences
Patents, Royalties, Other Intellectual Property: Co-inventor and patent royalty recipient of Anchored Multiplex PCR (AMP) technology

Miguel Rivera
Consulting or Advisory Role: Loxo Oncology, Asubio Pharmaceuticals (I)
Speakers' Bureau: Pfizer (I)
Research Funding: Advanced Cell Diagnostics, Affymetrix
Patents, Royalties, Other Intellectual Property: Patents with Affymetrix

Valentina Nardi
Stock and Other Ownership Interests: KSQ Therapeutics (I), The Navicor Group (I)
Consulting or Advisory Role: Thermo Fisher Scientific (I), Cell Signaling Technology (I)

Dora Dias-Santagata
No relationship to disclose

A. John Iafrate
Stock and Other Ownership Interests: Archer Biosciences
Consulting or Advisory Role: Debiopharm Group, Constellation Pharmaceuticals, Chugai Pharma, Roche
Research Funding: Blueprint Medicines
Patents, Royalties, Other Intellectual Property: ArcherDx exclusive license to AMP technology

Long P. Le
Stock and Other Ownership Interests: Archer Biosciences
Consulting or Advisory Role: Archer Biosciences
Patents, Royalties, Other Intellectual Property: Co-inventor and patent royalty recipient of AMP technology, which is licensed to ArcherDx
Travel, Accommodations, Expenses: Archer Biosciences

Jochen K. Lennerz
No relationship to disclose

ACKNOWLEDGMENT

We thank the entire clinical team of the Center for Integrated Diagnostics. We also thank Julie Batten, Hayley Robinson, Yi Cao, Caitlin E. Finn, and Amelia N. Raymond for expert technical assistance. Furthermore, we thank M.R. Toups; M. Boswell; D. Borger, PhD; J. Steinestel, MD; A. Stenzinger, MD; C. Wang, MD; J. Baron, MD; V. Klepeis, MD, PhD; D. Sgroi, MD; and D. Louis, MD, for thoughtful discussions.

REFERENCES

1. Haber DA, Gray NS, Baselga J: The evolving war on cancer. Cell 145:19-24, 2011
2. Sobel ME, Bagg A, Caliendo AM, et al: The evolution of molecular genetic pathology: Advancing 20th-century diagnostic methods into potent tools for the new millennium. J Mol Diagn 10:480-483, 2008
3. Goodwin S, McPherson JD, McCombie WR: Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet 17:333-351, 2016
4. Buermans HP, den Dunnen JT: Next generation sequencing technology: Advances and applications. Biochim Biophys Acta 1842:1932-1941, 2014
5. Hagemann IS, O'Neill PK, Erill I, et al: Diagnostic yield of targeted next generation sequencing in various cancer types: An information-theoretic approach. Cancer Genet 208:441-447, 2015
6. Roy S, LaFramboise WA, Nikiforov YE, et al: Next-generation sequencing informatics: Challenges and strategies for implementation in a clinical environment. Arch Pathol Lab Med 140:958-975, 2016
7. Zehir A, Benayed R, Shah RH, et al: Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med 23:703-713, 2017
8. Sharma MK, Phillips J, Agarwal S, et al: Clinical genomicist workstation. AMIA Jt Summits Transl Sci Proc 2013:156-157, 2013
9. Yohe SL, Carter AB, Pfeifer JD, et al: Standards for clinical grade genomic databases. Arch Pathol Lab Med 139:1400-1412, 2015
10. Zehnbauer BA, Buchman TG: Precision diagnosis is a team sport. J Mol Diagn 18:1-2, 2016
11. Aziz N, Zhao Q, Bry L, et al: College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. Arch Pathol Lab Med 139:481-493, 2015
12. Lai Z, Markovets A, Ahdesmaki M, et al: VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res 44:e108, 2016
13. Liu X, Han S, Wang Z, et al: Variant callers for next-generation sequencing data: A comparison study. PLoS One 8:e75619, 2013
14. Kircher M, Witten DM, Jain P, et al: A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310-315, 2014
15. Krøigård AB, Thomassen M, Lænkholm AV, et al: Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. PLoS One 11:e0151664, 2016
16. Dienstmann R, Dong F, Borger D, et al: Standardized decision support in next generation sequencing reports of somatic cancer variants. Mol Oncol 8:859-873, 2014
17. Zutter MM, Bloom KJ, Cheng L, et al: The cancer genomics resource list 2014. Arch Pathol Lab Med 139:989-1008, 2015
18. Patel NM, Michelini VV, Snell JM, et al: Enhancing next-generation sequencing-guided cancer care through cognitive computing. Oncologist 2017-0170, 2017
19. Castaneda C, Nalley K, Mannion C, et al: Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinforma 5:4, 2015
20. Baron JM, Dighe AS, Arnaout R, et al: The 2013 symposium on pathology data integration and clinical decision support and the current state of field. J Pathol Inform 5:2, 2014
21. Appenzeller T: The scientists' apprentice. Science 357:16-17, 2017
22. Hutson M: AI glossary: Artificial intelligence, in so many words. Science 357:19, 2017
23. Musib M, Wang F, Tarselli MA, et al: Artificial intelligence in research. Science 357:28-30, 2017
24. El Naqa I: Perspectives on making big data analytics work for oncology. Methods 111:32-44, 2016

25. Krittanawong C: The rise of artificial intelligence and the uncertain future for physicians. Eur J Intern Med [epub ahead of print on June 23, 2017] https://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=28651747&dopt=Abstract
26. Ciompi F, Chung K, van Riel SJ, et al: Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Sci Rep 7:46479, 2017
27. Moravčík M, Schmid M, Burch N, et al: DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. Science 356:508-513, 2017
28. Rocha JC, Passalia FJ, Matos FD, et al: A method based on artificial intelligence to fully automatize the evaluation of bovine blastocyst images. Sci Rep 7:7659, 2017
29. Schrum J, Miikkulainen R: Solving multiple isolated, interleaved, and blended tasks through modular neuroevolution. Evol Comput 24:459-490, 2016
30. Lee M, Roos P, Sharma N, et al: Systematic computational identification of variants that activate exonic and intronic cryptic splice sites. Am J Hum Genet 100:751-765, 2017
31. Ghanat Bari M, Ung CY, Zhang C, et al: Machine learning-assisted network inference approach to identify a new class of genes that coordinate the functionality of cancer networks. Sci Rep 7:6993, 2017
32. Arnott D: Cognitive biases and decision support systems development: A design science approach. Inf Syst J 16:55-78, 2006
33. Dias-Santagata D, Akhavanfard S, David SS, et al: Rapid targeted mutational analysis of human tumours: A clinical platform to guide personalized cancer medicine. EMBO Mol Med 2:146-158, 2010
34. Sequist LV, Heist RS, Shaw AT, et al: Implementing multiplexed genotyping of non-small-cell lung cancers into routine clinical practice. Ann Oncol 22:2616-2624, 2011
35. Zheng Z, Liebers M, Zhelyazkova B, et al: Anchored multiplex PCR for targeted next-generation sequencing. Nat Med 20:1479-1484, 2014
36. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760, 2009
37. Cibulskis K, Lawrence MS, Carter SL, et al: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31:213-219, 2013
38. Ramos AH, Lichtenstein L, Gupta M, et al: Oncotator: Cancer variant annotation tool. Hum Mutat 36:E2423-E2429, 2015
39. McLaren W, Gil L, Hunt SE, et al: The Ensembl variant effect predictor. Genome Biol 17:122, 2016
40. Kohavi R: A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Jt Conf Artif Intell 2:1137-1143, 1995
41. Kohavi R, Provost F: Machine Learning. Glossary of Terms. Boston, MA, Kluwer Academic, 1998
42. Arlot S, Celisse A: A survey of cross-validation procedures for model selection. Stat Surv 4:40-79, 2010
43. Rodríguez JD, Pérez A, Lozano JA: Sensitivity analysis of kappa-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell 32:569-575, 2010
44. Saeys Y, Inza I, Larrañaga P: A review of feature selection techniques in bioinformatics. Bioinformatics 23:2507-2517, 2007
45. Pedregosa F, Varoquax G, Michel V, et al: Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825-2830, 2011
46. Cruz-Roa A, Gilmore H, Basavanhally A, et al: Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. Sci Rep 7:46450, 2017

47. Buggenthin F, Buettner F, Hoppe PS, et al: Prospective identification of hematopoietic lineage choice by deep learning. Nat Methods 14:403-406, 2017

48. Chang K, Bai HX, Zhou H, et al: Residual convolutional neural network for determination of IDH status in low- and high-grade gliomas from MR imaging. Clin Cancer Res [epub ahead of print on November 22, 2017]

49. Stanford: One hundred year study on artificial intelligence (AI100), 2016. https://ai100.stanford.edu