correlate survival prognosis or drug treatment response. We have published a study showing that machine learning can separate patients by expected survival better than existing methods.[3]

Of course, there is interest in applying machine learning to a wide range of medical applications—analyzing clinical notes, looking for biomarkers in the pattern of lab tests, and analyzing genomics data. There are many insights about basic biology and clinical medicine undiscovered in large databases. A current focus area is the analysis of cancer "omics" datasets. The genome, transcriptome, metabolome, and proteome provide rich data reflecting the biology of cancer cells. These datasets are complex not only in their volume, but in their temporal trajectory and in the makeup of the cells that contribute to these measurements. Machine learning methods will recognize the key patterns that are associated with prognosis and treatment response. Deep learning methods can process histone epigenetic datasets[4] and predict the behavior of histones in induced pluripotent and differentiated cells.[5] Ultimately, AI methods will untangle the complex relationships between transcription factors and gene expression. Learning systems can predict the sequence specificity of DNA- and RNA-binding proteins, protein–chemical interactions—and even the outcome of phase I/II trials.[6]

The initial demonstrations of the promise of machine learning methods come with challenges, among the most pressing are:

1. Developing methods for integrating heterogeneous datasets. Most methods for collecting data are biased and reliable discoveries/analyses often come from several lines of independent data. Learning systems must improve in their ability to work with incomplete and heterogeneous data sources.
2. Creating curated datasets of sufficient quality to produce trustworthy classifications. The availability of high-quality datasets for training is often limiting. Expert curators must spend many hours creating datasets to train high-quality classifiers. The availability of large corpora of labeled radiographs and histopathology slides spurred progress in these areas.
3. Incorporating prior knowledge into the learning models. One way to mitigate the absence of large high-quality datasets is to give the models a "head start" by encoding human knowledge into an initial model, and allowing the system to refine it. A Bayesian "prior" model requires less data to converge on an excellent classifier.
4. Ensuring that the output of learning systems can be explained to human decision makers. Critical to both regulatory approval and clinical implementation is the ability of complex algorithms to justify their output: what are the key features driving the decision to classify a sample as "malignant" or to recommend an unusual drug? Human decision-makers must understand the evidence used and how it influenced the output.

A revolution in AI is being felt everywhere, including medical research and clinical medicine. The ability to find patterns in huge and complex datasets has been demonstrated in imaging and speech—where humans already excel. The technologies are now developing in areas such as cancer genomics, where high performance will require a partnership between human expert decision makers and AI systems that can find hidden patterns within these large rich datasets.

CONFLICT OF INTEREST

The author declares no conflicts of interest.

1. <http://www.nydailynews.com/news/world/ibm-watson-proper-diagnosis-doctors-stumped-article-1.2741857>.
2. <https://research.googleblog.com/2014/11/a-picture-is-worth-thousand-coherent.html>.
3. Yu, K.H. *et al*. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**, 12223 (2016).
4. BioRXiv, preprint. <https://doi.org/10.1101/052118>.
5. BioRXiv, preprint. <https://doi.org/10.1101/091660>.
6. BioRXiv, preprint. <https://doi.org/10.1101/095653>.

Data Sharing, Clinical Trials, and Biomarkers in Precision Oncology: Challenges, Opportunities, and Programs at the Department of Veterans Affairs

LD Fiore[1], MT Brophy[1], RE Ferguson[1], C Shannon[1], SJ Turek[1], K Pierce-Murray[1], S Ajjarapu[1], GD Huang[1], CSE Lee[1] and PW Lavori[1]

Cancer genomic research reveals that a similar cancer clinical phenotype (e.g., non-small cell lung cancer) can arise from various mutations in tumor DNA. Thus, organ of origin is not a defini-

[1]VA Boston Healthcare System, Boston, Massachusetts, USA. Correspondence: LD Fiore (Louis.Fiore@va.gov)

doi:10.1002/cpt.660

tive classification. Further, targeted therapy for cancer patients (precision oncology) capitalizes on knowledge of individual patient mutational status to deliver treatment directed against the protein products of these mutations with the goal of reducing toxicity and enhancing efficacy relative to traditional nontargeted chemotherapy.

THE NEED FOR COLLABORATIVE DATA SHARING

Subclassification and study of cancer patients based on mutational status presents opportunities to learn the significance of genomic alterations (and their combinations) and to develop additional therapies (and combinations of therapies). However, the large number of mutations known to be important in cancer development and the presence of multiple mutations in any individual patient combines to create a great diversity in populations of patients with a specific tumor type. Shrager and Tenenbaum note that cancer is "in effect, a large number of rare diseases occupying a very high dimensional space with very few opportunities for action and observation in each subtype. To efficiently search a space of this nature, one needs to capture the learnings from as many patients and treatment experiments as possible in a continuously updated knowledge base."[1]

The stratification of cancer patients by mutational status and resultant decrease in the proportion of patients available for study enrollment presents serious problems for observational and interventional research. The low prevalence (1–2%) of many driver mutations in solid tumors precludes recruitment of sufficient numbers of subjects from traditionally sized research consortia and has led to new models of clinical research based on interinstitution collaboration, community outreach, and broad data sharing, with the goal of learning from every treatment encounter. This article reviews the challenges and opportunities of such collaborations from the perspective of the Department of Veterans Affairs Healthcare System, the largest integrated healthcare system in the United States.

Creating generalizable knowledge as well as informing current individual patient care is therefore enabled by "learning" from all available previous treatment experiences of every relevant patient in the entire system, aggregating data across many medical centers. Familiar challenges to this approach include technical issues, such as data element provenance, data quality, and database variability across institutions, ensuring patient protections in data sharing related to informed consent, privacy/confidentiality, and Health Insurance Portability and Accountability Act (HIPAA) authorization, scalability and sustainability of aggregated databases, and cultural and financial barriers to data sharing in a research community.

The ability to commoditize healthcare data has created opportunities for new consortium models that enable data sharing and patient access (for clinical trials). Data sharing is an important element of the collaborations exemplified by ORIEN,[2] MED-C,[3] TAPUR,[4] and APOLLO[5] (see **Table 1**). The National Cancer Institute (NCI) has created the Genomic Data Commons, a unified data repository[6] that enables data sharing across these and other cancer genomic studies, in support of precision medicine. The repository houses clinical health record and genomic data (FASTQ file format) and complements the Cancer Imaging Archive,[7] another NCI-sponsored repository that contains radiographic and pathology images for cancer patients.

The Department of Veterans Affairs (VA) has begun to move consented and HIPAA-authorized patient data from the VA electronic medical records to these NCI data repositories for subsequent sharing with the research community (see **Table 1**). This approach replaces the need for cancer patients to sign multiple forms consenting to data sharing with a single broad consent that satisfies the requirements of the various project Institutional Review Boards (IRBs). It also provides data collection and curation that meets requirements of the individual project aims. The "Big Data Scientist Training and Education Program" is a joint effort by the VA and the NCI to provide these data to early career scientists to develop analytics and other tools in support of clinical and research objectives.[8]

CLINICAL TRIALS

Clinical trialists confront obstacles unique to precision oncology. As discussed above, traditional recruitment approaches from single or limited groups of institutions do not provide sufficient numbers of eligible study subjects to fulfill inclusion criteria with specific tumor–mutation combination requirements. Furthermore, third-party payers seldom reimburse for mutational analysis required for patient screening prior to entry into research, thus shifting to the research enterprise the cost of screening large numbers of patients, of whom only a small fraction will be found eligible. The lack of data on the mutational status of patients is the major bottleneck for clinical trial execution and slows progress in precision oncology. In the era of precision oncology treatment, the new standard of care requires clinical reimbursement for expanded panel testing in cancer patients, with subsequent recruitment in clinical trials when appropriate for the individual patient. That standard then leads to sharing and reuse of patient data for clinical trials and observational research.

Major efforts underway to solve these and related problems are exemplified by the MED-C and ORIEN initiatives. The MED-C Program offers insurance coverage of testing for institutions and patients who agree to contribute clinical data to the N1 registry for subsequent analysis. The Oncology Research Information Exchange Network (ORIEN) is a research

Table 1 Sharing with the research community

Group	Leading institutions	Consortium activities (beyond data sharing)
ORIEN[2]	The Moffitt and The Ohio State University Comprehensive Cancer Centers	Biobank and investigational drugs procurement
MED-C[3]	Molecular Evidence Development Consortium	Reimbursement for tumor sequencing
TAPUR[4]	The American Society of Clinical Oncology	Availability of targeted therapy off label
APOLLO[5]	National Cancer Institute, Department of Veterans Affairs and the Department of Defense	Tissue and data for biomarker development (proteogenomics)

collaboration founded by the Moffitt Cancer Center in Tampa and the Ohio State University Comprehensive Cancer Center to match patients to targeted treatments and promote data-sharing activities. The group engages industry partners on sponsored projects across the clinical trials continuum and serves as a broker between research and healthcare communities. These and other programs (particularly NCI-sponsored clinical trial consortia) that foster collaboration between clinical care and research communities represent new models to advance precision oncology.

Membership of VA sites in NCI Consortia and programs such as MED-C, ORIEN, and TAPUR makes data sharing and clinical trial participation opportunities available for Veterans at participating VA Medical Centers but leaves behind patients at facilities that lack oncology research programs and infrastructure. This structural problem is not unique to the VA, as opportunities for patients to participate in cancer clinical trials are similarly limited in communities not located near cancer centers.

If participation in clinical trials is considered a new standard of care, then this phenomenon exposes a new and important access disparity in a healthcare system. The Distributed Enrollment Program under development at the VA Cooperative Studies Program presents a model to "move clinical trials to patients" remote from VA cancer centers by seeking pre-approval of cancer protocols by a central IRB, reduction of research-specific training requirements imposed on participating clinicians, and centralized trial management (such as data collection and submission). These enhancements are engineered to deliver "just in time" research opportunities for patients while preserving patient safety and data integrity. Indeed, much of the administrative overhead built into the clinical research apparatus in the name of "quality assurance," while highly appropriate for research designed primarily to benefit the broader community (such as registration studies for drugs where effective alternatives exist), may have reduced relevance in a precision oncology setting, where the patient's motivation to participate is first and foremost to obtain study drug, whether or not alternatives exist.

BIOMARKER DISCOVERY AND VALIDATION

By its nature, precision oncology is a biomarker-driven field critically dependent on acquisition of clinical biosamples and electronic medical record data for discovery and validation with lack of access to both resources as a central limitation to the pace of discovery. Collaborations to cost-share for data and tissue procurement between biotech, pharma, and healthcare systems, in a "precompetitive fashion," are emerging.[4] To reach full potential, such collaborations require more complete integration within healthcare systems, as exemplified in the APOLLO Program, a partnership between the NCI, DoD, and VA. In APOLLO, tumor tissue is made available for biomarker analysis (proteomics in this case) and the results transmitted back to healthcare providers if they are determined to offer incremental value to patient care, beyond mutational analysis. While observational data is useful in this regard, randomization accelerates learning, and results in more certain knowledge. For example, demonstration that patients randomized to genomic augmented with proteomic analysis had superior treatment outcomes to those randomized to genomic analysis alone supports a compelling argument to adopt this new biomarker of response to targeted therapy. Problematically, the introduction of randomization (to biomarkers) into clinical care using traditional clinical trial methods is cost-prohibitive. The Department of Veterans Affairs continues to make progress in this area through the Point-of-Care research[9] and Precision Oncology[10] Programs whereby patients are randomized to minimal risk alternatives with relaxed regulatory requirements appropriate with the degree of risk (risk-based monitoring). Data generated from these embedded studies are derived exclusively from the EHR (real-world evidence) and require FDA acceptance if used for registration of a new companion diagnostic.

DISCLAIMER

The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government

1. Shrager, J. & Tenenbaum, J.M. Rapid learning for precision oncology. *Nat. Rev. Clin. Oncol.* **11**, 109–118 (2014).
2. <http://oriencancer.org/>. Accessed 25 January 2017.
3. <https://med-c.org/>. Accessed 25 January 2017.
4. <http://www.tapur.org/>. Accessed 25 January 2017.
5. APOLLO is discussed elsewhere in this edition.
6. <https://gdc.cancer.gov/>. Accessed 25 January 2017.
7. <http://www.cancerimagingarchive.net/>. Accessed 25 January 2017.

8. <https://www.va.gov/oaa/specialfellows/programs/sf_bdstep.asp?p>. Accessed 25 January 2017.
9. Fiore, L. *et al*. A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen. *Clin. Trials* **8**, 183–195 (2011).
10. Fiore, L.D. *et al*. The VA Point-of-Care Precision Oncology Program: balancing access with rapid learning in molecular cancer medicine. *Biomark. Cancer* **8**, 9–16 (2016).

Collaborating to Compete: Blood Profiling Atlas in Cancer (BloodPAC) Consortium

RL Grossman[1], B Abel[2], S Angiuoli[3], JC Barrett[4], D Bassett[5], K Bramlett[6], GM Blumenthal[7], A Carlsson[8], R Cortese[9], J DiGiovanna[9], B Davis-Dusenbery[9], R Dittamore[10], DA Eberhard[2], P Febbo[2], M Fitzsimons[1], Z Flamig[1], J Godsey[30], J Goswami[31], A Gruen[9], F Ortuño[1], J Han[2], D Hayes[11], J Hicks[8], D Holloway[9], D Hovelson[11], J Johnson[4], H Juhl[12], R Kalamegham[13], R Kamal[14], Q Kang[11], GJ Kelloff[15], M Klozenbuecher[14], A Kolatkar[8], P Kuhn[8], K Langone[2], R Leary[16], P Loverso[3], H Manmathan[9], A-M Martin[17], J Martini[23], D Miller[1], M Mitchell[18], T Morgan[11], R Mulpuri[19], T Nguyen[1], G Otto[20], A Pathak[21], E Peters[32], R Philip[21], E Posadas[22,28], D Reese[19], MG Reese[14], D Robinson[16], A Dei Rossi[2], H Sakul[23], J Schageman[6], S Singh[20], HI Scher[18], K Schmitt[1], A Silvestro[16], J Simmons[3], T Simmons[1], J Sislow[1], A Talasaz[24], P Tang[1], M Tewari[11], S Tomlins[11], H Toukhy[24], HR Tseng[22,29], M Tuck[11], A Tzou[21], J Vinson[18], Y Wang[10], W Wells[25], A Welsh[20], J Wilbanks[26], J Wolf[19], L Young[20], JSH Lee[15] and LC Leiman[27]

The cancer community understands the value of blood profiling measurements in assessing and monitoring cancer. We describe an effort among academic, government, biotechnology, diagnostic, and pharmaceutical companies called the Blood Profiling Atlas in Cancer (BloodPAC) Project. BloodPAC will aggregate, make freely available, and harmonize for further analyses, raw datasets, relevant associated clinical data (e.g., clinical diagnosis, treatment history, and outcomes), and sample preparation and handling protocols to accelerate the development of blood profiling assays.

[1]Center for Data Intensive Science, University of Chicago, Chicago, Illinois, USA; [2]Genomic Health, Redwood City, California, USA; [3]Personal Genome Diagnostics, Baltimore, Maryland, USA; [4]AstraZeneca, Waltham, Massachusetts, USA; [5]Celgene, Seattle, Washington, USA; [6]Thermo Fisher Scientific, Austin, Texas, USA; [7]Center for Drug Evaluation and Research, Food and Drug Administration, Silver Springs, Maryland, USA; [8]Department of Molecular and Medical Pharmacology, Crump Institute for Molecular Imaging, University of California, Los Angeles, California, USA; [9]Seven Bridges, Cambridge, Massachusetts, USA; [10]Epic Research and Diagnostics, San Diego, California, USA; [11]University of Michigan, Ann Arbor, Michigan, USA; [12]Indivumed GmbH, Hamburg, Germany; [13]Genentech, Washington, District of Columbia, USA; [14]Omicia, Oakland, California, USA; [15]Office of the Director, National Cancer Institute, Bethesda, Maryland, USA; [16]Novartis Institute for Biomedical Research, Cambridge, Massachusetts, USA; [17]Novartis Pharmaceuticals, East Hanover, New Jersey, USA; [18]Memorial Sloan Kettering Cancer Center, New York, New York, USA; [19]Provista Diagnostics Inc., New York, New York, USA; [20]Foundation Medicine, Cambridge, Massachusetts, USA; [21]Center for Device and Radiological Health, Food and Drug Administration, Silver Springs, Maryland, USA; [22]CytoLumina, Inc., Los Angeles, California, USA; [23]Pfizer, San Diego, California, USA; [24]Guardant Health, Inc., Redwood City, California, USA; [25]Open Commons Consortium, Chicago, Illinois, USA; [26]Sage Bionetworks, Seattle, Washington, USA; [27]BloodPAC, Chicago, Illinois, USA; [28]Cedar-Sinai Medical Center, Los Angeles, California, USA; [29]Crump Institute for Molecular Imaging, University of California, Los Angeles, California, USA; [30]Thermo Fisher Scientific, Waltham, Massachusetts, USA; [31]Thermo Fisher Scientific, Carlsbad, California, USA; [32]Genentech, South San Francisco, California, USA. Correspondence: RL Grossman (robert.grossman@uchicago.edu)

doi:10.1002/cpt.666