Insights to Accelerate Progress in Assessment of Drug Effect – Including Cancer Imaging Technology

Robert M Califf MD
Advisor, Verily Life Sciences
Vice Chancellor for Health Data Science, Duke University
November 11th, 2019
Conflicts of interest

● Employment
  ● Duke University
  ● Verily Life Sciences

● Corporate Board
  ● Cytokinetics

● Consulting
  ● Astra Zeneca
  ● Merck
  ● Boehringer Ingelheim
  ● Amgen
  ● Biogen
  ● Genentech/Roche
  ● Lilly
Verily partners closely with Google teams on commercial tools and applications across the healthcare vertical.
Generating Evidence to Inform Decisions

1. FDA Critical Path
2. NIH Roadmap
3. Data Standards
4. Network Information
5. Empirical Ethics
6. Priorities and Processes
7. Inclusiveness
8. Use for Feedback on Priorities
9. Conflict of Interest Management
10. Evaluation of Speed and Fluency
11. Pay for Performance
12. Transparency to Consumers

Early Translational Steps

Discovery Science

Outcomes

Performance Measures

Measurement and Education

Clinical Trials

Clinical Practice Guidelines

Generates Evidence to Inform Decisions
Key Points

- Given systematic imaging data and well curated clinical data at scale
  - ML/AI assistance will improve reliability of the measurement
  - The information content will improve because of the multi-dimensionality of the imaging information
  - Combining imaging information with other clinical information will produce better estimates than clinical information alone
- Cancer patients need us to figure out how to share data
Health Status and Determinants

Figure 1. Life expectancy at birth, by sex and race and Hispanic origin: United States, 2007–2017

NOTES: Some states reported multiple-race data. The multiple-race data for these states were bridged to the four single-race categories for comparability across the trend. Life expectancy estimates for 2017 use preliminary Medicare data. For more information, see Appendix B, Life expectancy. See data table for Figure 1. SOURCE: NCHS, National Vital Statistics System (NVSS), Mortality. Excel and PowerPoint: https://www.cdc.gov/nchs/hus/contents2018.htm#Figure_001
LIFE EXPECTANCY AT BIRTH BY COUNTY, 2014

- Counties in South Dakota and North Dakota had the lowest life expectancy, and counties along the lower half of the Mississippi, in eastern Kentucky, and southwestern West Virginia also had very low life expectancy compared with the rest of the country. Counties in central Colorado had the highest life expectancies.

CHANGE IN LIFE EXPECTANCY AT BIRTH BY COUNTY, 1980 TO 2014

- Compared with the national average, counties in central Colorado, Alaska, and along both coasts experienced larger increases in life expectancy between 1980 and 2014, while some southern counties in states stretching from Oklahoma to West Virginia saw little, if any, improvement over this same period.

Life expectancy at birth (years) in 18 high income countries for women and men during 2010-16 and 1990-2015.

Jessica Y Ho, and Arun S Hendi BMJ 2018;362:bmj.k2562
Countries with universal healthcare

"Universal healthcare can't possibly work!"

Cheaper and more effective

More expensive, less effective

Failed attempts to institute universal healthcare system in US
Number of deaths among persons aged <80 years from the five leading causes of death, by urban-rural county classification — National Vital Statistics System, United States, 2017
Annual percent change in potentially excess deaths* among persons aged <80 years from the five leading causes of death, by urban-rural county classification — National Vital Statistics System. United States. 2010–2017

Cancer

Heart disease

Unintentional injury

Chronic lower respiratory disease

Stroke

Cause of death

Diagram showing the annual percent change in potentially excess deaths among persons aged <80 years from the five leading causes of death, by urban-rural county classification for the years 2010–2017.
Percentage of deaths that were potentially excess* among persons aged <80 years from the five leading causes of death, by urban-rural county classification — National Vital Statistics System, United States, 2017
Innovations jointly deployed by Google + Verily

Healthy

Diseased

Hemorrhages

No DR  Mild DR  Moderate DR  Severe DR  Proliferative DR
Performance in tumor localization - Camelyon16 challenge data set

Tumor localization score (FROC):
- Single pathologist: 0.73*
- Camelyon16 winner: 0.81
- Google AI algorithm: 0.91

The algorithm also generalizes to data from other clinics and scanners

* unlimited time (30h), but 0 false positives

Slide level AUC:
- Single pathologist: 96.6%
- Google AI Algorithm: 99.3%
The Deep Learning System is better at quantitating Gleason patterns than board-certified pathologists.

- DLS shows significantly improved quantitation relative to specialist-adjudicated reference standard.
- Particularly in GG2-3 slides:
  - %GP4 can change overall Grade Group
  - 5-10% increments of %GP4 are prognostic
- Limitation: The reference truth is still derived from visual estimation.

Nagpal et al, npj Digital Medicine, June 2019
Clinical Outcomes Analysis Consistent with Notion that Specialist Grades are More Prognostic than Generalists

<table>
<thead>
<tr>
<th></th>
<th>Hazard Ratio at Grade Group GG≥3 threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort-of-29 Pathologists</td>
<td>1.14 (0.53-1.84)</td>
</tr>
<tr>
<td>Deep Learning System</td>
<td>1.38 (0.63-2.12)</td>
</tr>
<tr>
<td>Specialist-defined reference standard</td>
<td>1.54 (0.90-2.40)</td>
</tr>
</tbody>
</table>

- Trend is encouraging
- Validation on larger datasets are needed to demonstrate statistical significance

Larger delta = Better stratification

n=320

Nagpal et al, npj Digital Medicine, June 2019
Lung Diagnosis from CT Outperformed Specialists

Abstract

With an estimated 160,000 deaths in 2018, lung cancer is the most common cause of cancer death in the United States\(^1\). Lung cancer screening using low-dose computed tomography has been shown to reduce mortality by 20–43% and is now included in US screening guidelines\(^2\).\(^3\).\(^4\).\(^5\).\(^6\). \(^7\).\(^8\).\(^9\).\(^10\). Existing challenges include inter-grader variability and high false-positive and false-negative rates\(^7\).\(^8\).\(^9\).\(^10\). We propose a deep learning algorithm that uses a patient’s current and prior computed tomography volumes to predict the risk of lung cancer. Our model achieves a state-of-the-art performance (94.4% area under the curve) on 6,716 National Lung Cancer Screening Trial cases, and performs similarly on an independent clinical validation set of 1,139 cases. We conducted two reader studies. When prior computed tomography imaging was not available, our model outperformed all six radiologists with absolute reductions of \(11\%\) in false positives and \(3\%\) in false negatives. Where prior computed tomography imaging was available, the model performance was on-par with the same radiologists. This creates an opportunity to optimize the screening process via computer assistance and automation. While the vast majority of patients remain unscreened, we show the potential for deep learning models to increase the accuracy, consistency and adoption of lung cancer screening worldwide.

Fig. 2: Results from the reader study—lung cancer screening on a single CT volume.
FDA Analysis of Discordance between Investigator and Independent Review Committee (relevant article in press)

<table>
<thead>
<tr>
<th>Best Overall Response (IRC)</th>
<th>CR</th>
<th>PR</th>
<th>SD</th>
<th>PD</th>
<th>NE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR</td>
<td>305</td>
<td>250</td>
<td>116</td>
<td>14</td>
<td>25</td>
<td>710</td>
</tr>
<tr>
<td>PR</td>
<td>240</td>
<td>3366</td>
<td>908</td>
<td>126</td>
<td>33</td>
<td>4673</td>
</tr>
<tr>
<td>SD</td>
<td>64</td>
<td>849</td>
<td>3522</td>
<td>668</td>
<td>74</td>
<td>5177</td>
</tr>
<tr>
<td>PD</td>
<td>9</td>
<td>58</td>
<td>507</td>
<td>1890</td>
<td>134</td>
<td>2598</td>
</tr>
<tr>
<td>NE</td>
<td>1</td>
<td>13</td>
<td>31</td>
<td>49</td>
<td>425</td>
<td>519</td>
</tr>
<tr>
<td>Total</td>
<td>619</td>
<td>4536</td>
<td>5084</td>
<td>2747</td>
<td>691</td>
<td>13677</td>
</tr>
</tbody>
</table>
## History of Digital Disruption

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kodak</td>
<td>Photographic Film &amp; Paper</td>
<td>Digital Photography</td>
<td>NO</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Barnes &amp; Noble</td>
<td>Selling Books from Stores</td>
<td>Online Book Orders</td>
<td>NO</td>
<td>X</td>
<td>Amazon.com</td>
</tr>
<tr>
<td>Chemical Bank</td>
<td>Lending Money</td>
<td>ATMs and Online Banking</td>
<td><strong>YES</strong></td>
<td></td>
<td>All Modern Banks</td>
</tr>
<tr>
<td>Blockbuster Video</td>
<td>Video Rental</td>
<td>Digital Streaming</td>
<td>NO</td>
<td>X</td>
<td>Netflix</td>
</tr>
<tr>
<td>Fee-for-Service Health Care</td>
<td>Value-based, Digitally Enabled Medicine</td>
<td>NO</td>
<td>?</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sell more drugs at higher prices</td>
<td>Value based reimbursement</td>
<td>NO</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Qualities of the New Data Environment

• **Volume**
  — New methods of data storage allow access to huge amounts of data

• **Ubiquity/Liquidity**
  — Data are available anywhere across geography, social and economic classes

• **Latency**
  — There is no delay in access to data inherent in the technology

• **Analysis**
  — Data, information, knowledge, wisdom continuum is being shifted to the right
“To learn the truth, we must put all the parts together.”
16.3M results in 0.57 second
Phenotype stack

- Integration across scales requires new data, new tools, new taxonomy
- Unifying metadata: **small molecules, biophysical stimuli**
- Breadth vs Depth
- Co-clinical modeling
Product of the Biomarker Working Group charged by the FDA-NIH Joint Leadership Council to develop a glossary of harmonized terminology for biomarkers and endpoints

(*Today’s talk will focus on biomarker-related terms)

Surrogate Endpoint Validation

Correlative versus trial-level surrogate endpoint


Reasons for failure of candidate surrogate endpoints (Figure 1 from Fleming & DeMets)

A. The candidate surrogate is not in the causal pathway of the disease process.

B. Of several causal pathways of disease, the intervention affects only the pathway mediated through the candidate surrogate.

C. The candidate surrogate is not in the pathway of the intervention’s effect or is insensitive to its effect.

D. The intervention has mechanisms of action independent of the disease process.
**BEST Definition of Biomarker**

A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions. Molecular, histologic, radiographic, or physiologic characteristics are types of biomarkers. A biomarker is not an assessment of how an individual feels, functions, or survives.

- Expand to measurement types beyond biochemical
- Not specific to therapeutics development
- Distinguished from *clinical outcomes*
Validation of a Biomarker Test

- **Analytical validation** - Establishing that the performance characteristics of the test are acceptable in terms of its sensitivity, specificity, accuracy, precision, as applicable.
  - Technical performance
  - Says nothing about clinical correlations
  - Poor analytical validation may impede clinical validation

- **Clinical validation** - Establishing that the test, acceptably identifies, measures, or predicts the concept of interest (i.e., aspect of an individual’s clinical, biological, physical, or functional state, or experience).
  - Establish clinical associations
  - Many statistically significant p-values in published literature
  - Not guaranteed to be useful

- **Fit-for purpose validation**
  - Qualification (regulatory mechanism to establish suitable for use in medical product development)
  - Clinical utility determination (favorable benefit-to-risk for clinical use)
BEST Biomarker Categories

- Susceptibility/risk biomarker
- Diagnostic biomarker
- Prognostic biomarker
- Predictive biomarker
- Monitoring biomarker
- Pharmacodynamic/response biomarker
- Safety biomarker
Prognostic versus Predictive
Importance of control groups

Prognostic but not predictive

No survival benefit from new treatment

(M = biomarker)

Prognostic and predictive

New treatment for all or for M+ only
Surrogate endpoint – An endpoint (usually a biomarker or a composite of a biomarker and clinical endpoint) that is used in clinical trials as a substitute for a direct measure of how a patient feels, functions, or survives (clinical outcome). A surrogate endpoint does not measure the clinical benefit of primary interest in and of itself, but rather is expected to predict that clinical benefit or harm based on epidemiologic, therapeutic, pathophysiologic, or other scientific evidence.

(Extensions to epidemiologic settings are possible, but are not currently addressed in the BEST glossary.)
Surrogate Endpoint Levels of Evidence

From a U.S. regulatory standpoint, surrogate endpoints and potential surrogate endpoints can be characterized by the level of clinical validation.

• **Validated** – can be used for full approval of medical products

• **Reasonably likely** – sometimes can be used for accelerated approval

• **Candidate surrogate endpoint** – evidence still accumulating
Surrogate Endpoint Levels of Evidence

**Validated surrogate endpoint** - An endpoint supported by a *clear mechanistic rationale* and *clinical data providing strong evidence* that an effect on the surrogate endpoint *predicts a clinical benefit*. Therefore, it can be *used to support traditional approval* without the need for additional efficacy information (“replacement endpoint”).

- HIV-RNA reduction is a validated surrogate endpoint for human immunodeficiency virus (HIV) clinical disease control and has been used for the basis for approval of drugs intended to treat HIV.
Surrogate Endpoint Levels of Evidence

**Reasonably likely surrogate endpoint** - An endpoint supported by clear mechanistic and/or epidemiologic rationale but insufficient clinical data to show that it is a validated surrogate endpoint. Such endpoints can be used for accelerated approval for drugs or expedited access for medical devices.

- Decrease in iron stores for patients with iron overload caused by thalassemia has been considered reasonably likely to predict a decrease in transfusion-related adverse events caused by iron overload in the body and has supported accelerated approval of drugs to treat non-transfusion-dependent thalassemia (NTDT).
Surrogate Endpoint Levels of Evidence

**Candidate surrogate endpoint** - An endpoint still under evaluation for its ability to predict clinical benefit.

- Many examples
- Very few biomarkers achieve validated surrogate status
Surrogate Endpoint Validation

Correlative versus trial-level surrogate endpoint

• Prentice criteria
  – Surrogate endpoint must be a correlate of the true (definitive) clinical outcome
  – Surrogate must fully capture the net effect of treatment on the true clinical outcome
    • Rarely holds
      • Might hold approximately for some treatments but not others

• If surrogacy holds at all, it might depend on
  – Particular true clinical outcome measure
  – Method of measuring the surrogate
  – Class of drug (or intervention)
  – Patient population
Surrogate Endpoint Validation
Correlative versus trial-level surrogate endpoint

“A correlate does not a surrogate make.”

• Examples of failed surrogate candidates:
  ventricular arrhythmias for cardiovascular-related mortality, prostate biopsy for prostate cancer mortality, tumor response for survival

Setting that provides the greatest potential for the surrogate endpoint to be valid (Figure 2 from Fleming & DeMets)
Figure 1 Overview of the imaging biomarker roadmap

Technical (assay) validation
Imaging biomarker evaluated in vitro, in animals and in humans

Biological and clinical validation
Imaging biomarker is a reliable measure used to test hypotheses in clinical cancer research

Cost effectiveness
Imaging biomarker routinely used in the management of patients with cancer within the healthcare system

O’Connor, J. P. B. et al. (2016) Imaging biomarker roadmap for cancer studies
Association Between FDA Label Restriction and Immunotherapy and Chemotherapy Use in Bladder Cancer

JAMA. 2019;322(12):1209-1211.
Key Points

● Given systematic imaging data and well curated clinical data at scale
  ● ML/AI assistance will improve reliability of the measurement
  ● The information content will improve because of the multi-dimensionality of the imaging information
  ● Combining imaging information with other clinical information will produce better estimates than clinical information alone

● Cancer patients need us to figure out how to share data

#FDAPDSsymp | #AlinHealth