Data Science Opportunities for the NCI

Interim Recommendations

National Cancer Advisory Board
Ad hoc Working Group on Data Science

August 14, 2018
New WG formed by Dr. Sharpless/NCAB

Dec-April
Co-Chairs identified
Members finalized/confirmed
Possible priority areas identified

F2F meeting at NCI
Charge to WG
Discussion of priority areas for data science
Brainstorming
Consensus and identification of longer-term priority areas
Prioritization and development of short-term targeted areas

Virtual “pre-meeting”
Introductions
Background

Leapfrog for data sharing
Terminology harmonization
Training
Challenges/prizes

June-July
Subgroup meetings to flesh out and develop initial recommendation areas

Full WG virtual meeting
Discussion of draft recommendations
Initial Recommendation Areas

1. Investments to leapfrog data sharing for high-value datasets
2. Harmonization of terminology between cancer research data and clinical care data
3. Support of data science training at the graduate level
4. Opportunities for funding challenges and prizes
Recommendation 1: Investments to leapfrog data sharing for high-value datasets

- Resources to support
  - Identification
  - Enrichment
  - Curation
  - Harmonization
  - Annotation
  - Publishing

- Examples of high-value datasets
  - Those fully collected and annotated but not yet shared in a public repository
  - Datasets that would be enhanced by additional data generation and/or collection (e.g., genomic datasets needing additional clinical annotation)

Subgroup members:
- John Carpten
- Warren Kibbe
- Mia Levy
- Vince Miller
- Charles Sawyers
- Nick Wagle
Recommendation 2: Harmonize terminologies between cancer research and clinical care

• Augment EHR data standards to further bridge clinical care and cancer research
• Fund research related to achieving near clinical trial grade data within traditional clinical care settings
• Identify and prioritize existing standards bodies and activities

Subgroup members:
• George Hripcsak
• Mimi Huizinga
• Warren Kibbe
• Michelle Le Beau
Benefits of harmonized terminologies

• Increase the utility and ease of incorporation/integration of clinical care data from EHRs into cancer research

• Enable more efficient research, better patient care, and better real-world evidence generation

• Enhance integration of the cancer and non-cancer research communities
Recommendation 3: Increase the number of training programs and trainees in cancer data science

- Dedicate a specific T32 training program in cancer data science
- Contribute to existing NIH training programs
  - NLM T15 training programs
  - NIGMS Medical Scientist Training program
- Develop a short-term training program for clinicians and biological scientists

Subgroup members:
- Regina Barzilay
- Amanda Haddock
- Rebecca Jacobson
- Anne-Marie Meyer
- Sylvia Plevritis
- Kim Sabelko
Recommendation 4: Sponsor a series of data science challenges

- Potential challenge topics (~4-8 per year)
  - Drug response prediction
  - Discovery of multi-omic prognostic biomarkers
  - De-convolution of heterogenous tumors
  - Cancer diagnosis, grading, and staging
  - Facility of data access and integration from the ethical, legal, and social implications standpoint

- Consider beginning with an “idea challenge” to identify the appropriate challenge topic/task/question

Subgroup members:
  - Regina Barzilay
  - Amanda Haddock
  - Michelle Le Beau
  - Lincoln Stein
Benefits of data science challenges

• Spur research in computational cancer biology and increase the availability of advanced analytic software to the broader research community
• Attract new talent to cancer research
• Validation and dissemination of state-of-the-art tools and technologies
• Demonstrates the inter-relationship between all the recommendations. Challenges require:
  • Openly shared datasets
  • Ability to work across harmonized datasets
  • Participants with appropriate skillsets and expertise