PRECISION MEDICINE

Do we need Artificial Intelligence for Precision Medicine?

Prof. Robert Clarke, PhD, DSc

THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA

What is Artificial Intelligence (AI)

"Ability of a computer or computer-controlled robot to perform tasks commonly associated with intelligent beings" (*Encyclopedia Britannica*)

In precision medicine: the use of computer algorithms (often using machine learning) and workflows to learn how to predict a future state, e.g., disease risk, outcome, optimal therapy, etc.

Prognostic and Predictive Biomarkers

Predictions: use data at (or up to) one point in time to estimate the likely state of the system at some future time

Biomarkers: usually identified at the population level and used to predict an individual's disease risk or other outcome (future state) relative to the population

Prognostic biomarkers: personalized prediction of future disease state

Predictive biomarkers: personalized prediction of the intervention(s) that will produce the optimal clinical outcome

Prediction and Precision

Prediction: estimate (often expressed as a probability and/or summary statistic) of the future risk of an individual developing a disease, a specific clinical outcome for a patient (e.g., within a specified period of time), the optimal intervention for a patient, a molecular target for therapy, etc.

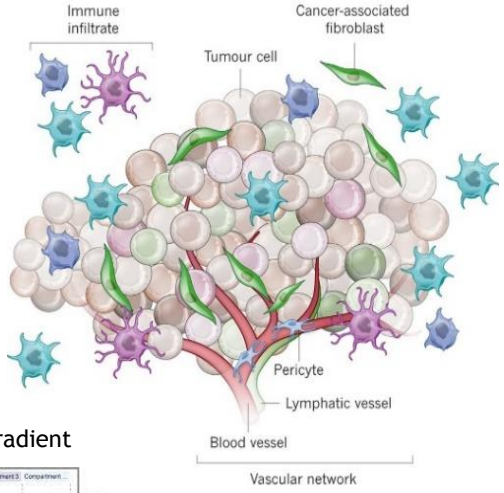
Precision: degree of sensitivity (true positives) and specificity (true negatives) of a prediction algorithm

Data Sources for Precision Medicine

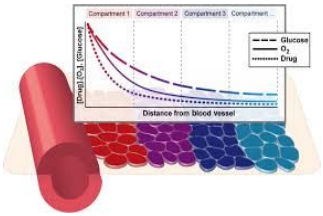
- Case history, disease status, etc.
- Specimen histopathology
- Validated individual biomarkers, e.g., HER2 (breast cancer)
- Validated complex biomarkers, e.g., PAM50 (breast cancer)
- Omics data for biomarker discovery/validation and target discovery for drug discovery/repurposing
  - Often genome, transcriptome, proteome data from patient samples

Data Sources: Heterogeneity

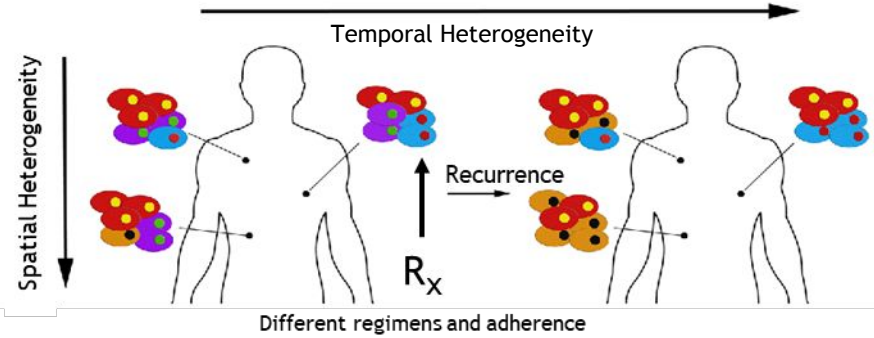
Intra-tumoral Heterogeneity



R_x Perfusion Gradient



Inter-tumoral Heterogeneity

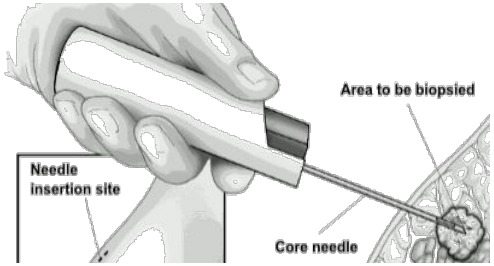


Patient Data Heterogeneity

- Age
- Socio-economics
- Race/ethnicity
- Comorbidities
- Treatment(s)
- Etc.

Data Sources: Heterogeneity

Sampling Bias



Sampling bias can affect the representation of tissue features

- Genetic or epigenetic heterogeneity (e.g., mutations, DNA methylation)
- Molecular heterogeneity (e.g., proteins, metabolites)
- Cellular heterogeneity (e.g., stromal, tumor, immune cells)
- Drug and nutrient perfusion heterogeneity
- Drug response heterogeneity

Addressing Heterogeneity

Data deconvolution (e.g., genetic, molecular, cellular heterogeneity)

- Supervised tools (several)
  - Deconvolution is supervised by external data
  - Knowledge of the number of cell types present, e.g., histopathology
  - Data for adequate supervision is often unavailable
- Unsupervised tools (few)
  - Deconvolution is done without reference to any external data
  - Required where data for algorithm supervision is unavailable
- Alternative: tissue microdissection and single cell sequencing
  - Currently limited transcriptome coverage (misses 50-75% of the transcriptome)
  - Lower coverage for the proteome and metabolome
  - Sampling bias (how many single cells capture the heterogeneity)

Data Properties: Dimensionality

Epidemiology

- 10,000 subjects
- Questionnaire with 100 questions
- 100-dimensional data with 10,000 samples

Transcriptome Assay

- 10,000 mRNAs in single cell RNAseq study
- 100 specimens
- 10,000-dimensional data with 100 samples

How well do we sample what's really present?

- ~30,000 genes
- ~50,000 RNA transcripts (all types)
- perhaps 80,000-400,000 different proteins
- >110,00 metabolites (HMDB 4.0)
- Likely many protein-protein, protein-metabolite, protein-DNA, and protein-RNA connections

Data Properties: Dimensionality

- Concentration of measure
  - Data are not evenly distributed in high dimensional data spaces
- Curse of dimensionality
  - Large search radius, number of calculations increase exponentially with dimensionality, algorithms allocate resources to irrelevant data regions, algorithms may converge on local solutions that are globally incorrect
- Multimodality (in biological systems)
  - More than one process, pathway, or phenotype is present
  - A signaling feature (e.g., gene, signaling module) may affect more than one component of a complex phenotype
- Confound of multimodality (complex phenotype)
  - Which component(s) of the profile uniquely define the phenotype of interest (e.g., biomarker study)
  - Which omics feature(s) truly reflect which phenotype component (e.g., mechanistic study to find new drug target)

Addressing Dimensionality

- Visualization
  - Examine all data by multidimensional scaling (e.g., PCA)
  - Validate retention of data structure after dimension reduction
- Reduce dimensionality to
  - Eliminate redundancy or uninformative data
  - Reduce noise
  - Ease the curse of dimensionality
  - Improve algorithm performance
- Reduce dimensionality by
  - Removing features (e.g., genes) lacking variable expression
  - Removing features not associated with a surrogate for outcome
  - Multiple t-testing (without correction for multiple comparisons)
  - Filters, e.g., fold regulation, abundance
  - Contribution to data variance (e.g., PCA)

PCA = principal component analysis

Addressing Dimensionality

- Biomarker studies can be viewed as pattern recognition problems
  - Goal is usually to find a pattern (one or more features) that when present/absent, high/low, etc. accurately and robustly predict specific phenotypes or outcomes (e.g., prognosis, treatment responsiveness)
  - Often the pattern is to be identified from within high dimensional data
- Support Vector Machines
  - Linear model for classification
  - Identifies the hyperplane that best separates data points
  - Largely unaffected by dimensionality
  - Can incorporate a recursive feature elimination process to find the smallest number of features needed to enable good classification
- New approaches continue to emerge

Study Design Goals

Need for, and approaches to, addressing heterogeneity and dimensionality are related to the study goal(s)

- Molecular Profiling (e.g., biomarker discovery)
  - **question:** what genes can define a specific phenotype?
  - **goal:** class prediction (identify class membership of an unknown sample)
  - **goal:** class discovery (identify new classes)
  - Molecular profile may or may not offer mechanistic insight(s)
- Mechanistic Studies (e.g., target discovery for drug discovery/ repurposing)
  - **question:** what actionable genes are responsible for a specific phenotype?
  - **goal:** gene selection

Summary: AI and Precision Medicine

AI makes working with complex and high dimensional data tractable, for example:

- Addressing heterogeneity
  - Data deconvolution (supervised or unsupervised) to learn the prevalence of cell types or different molecular features
- Addressing high dimensionality
  - Incorporating dimension independent tools into discovery workflows, e.g., SVM for classification
- Biomarker discovery (non-mechanistic)
  - Learning the most accurate and robust classifier
- Target discovery for drug discovery/repurposing (mechanistic)
  - Discovering actionable molecular targets

Thank You!

THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA