RE-ENGINEERING THE FUTURE OF HEALTH
WITH PREDICTIVE MODELS

New York Scientific Data Summit (NYSDS)
Grace C.Y. Peng, Ph.D.
October 22, 2020
Virtual
The Future of Health
• Real-time measurements
• Digital Twins
• Salutogenesis

Predictive Modeling
• Types
• Process
• Credible practice

Biomedical and Behavioral Research
Culture
• Modeling approaches
• Data landscape
• Heterogeneity and uncertainty

How do we get there?
• Integrate heterogeneous data
• Standardize processes for credible, reusable models
• **Next Gen Models** to predict and provide insight
THE FUTURE OF HEALTH
The Future of Health

Future Medicine: Dynamic data, Digital Health

Time

THE DIGITAL HEALTH REVOLUTION
Infographic by Paul Sonnier

NIH
National Institute of Biomedical Imaging and Bioengineering

Digital Twins

Digital Twins in Chronic Disease
A near-real-time linkage between physical and digital worlds

Digital twins exist at the nexus of physical engineering, data science, and machine learning, and their value translates directly to measurable business outcomes.*

*The Digital Twin: Compressing time-to-value for digital industrial companies, GE
Salutogenesis

Healthy State  ↔  Semi-Healthy State

Behavioral Intervention

Pharmacological Intervention

Unhealthy State

Acute Illness
PREDICTIVE MODELING
Interagency Modeling and Analysis Group
Biological Spatiotemporal Scales

Hunter and Borg, Nature 2003
Why Model?

- Infrastructure for systematically archiving and transferring knowledge
  - Prevent reinventing processes year after year
- Only way to predict outcomes not otherwise testable
- Drive scientific discovery → emergent properties
- Extend insight and understanding beyond the cognitive capabilities of the human mind
Model-Driven Science

Modeling & Simulation
→ Testable hypotheses
→ Predictions

Creativity
Intuition
Inspiration
New Theory

Hypothesis → Study Design → Experiment → Results → Reporting

Simulation

Populate knowledge base
Generate theories

Emergent Properties

OUTLIERS
Predictive models generate new hypotheses, and do not merely recapitulate the data that were used to build them. [2003-present: IMAG Multiscale Modeling Initiative]
High Performance Computing – the AI wave

Mechanistic Models to drive better simulations

Correlations: Observations and patterns

Community knowledge landscape

Diversity within community

HPC augmented hypothesis testing

HPC platforms

Repository and access

Distributed community effort

Multiscale integration

An 2010


Ten Simple Rules of Credible Practice

One of the first tasks of the Committee was to identify best practices to enhance credibility of modeling & simulation in healthcare. This activity started as a Committee discussion, where CPMS Task Teams have been tasked with generating a list of ten key elements or simple rules of credible practice (Committee Perspective). As the Committee discussions finalized, the group agreed on the necessity to reach out to the broader population of stakeholders. In result, the Committee launched a public survey to establish the...
Ten Not So Simple Rules for Model Credibility

1. Define context clearly
2. Use appropriate data
3. Evaluate within context
4. List limitations explicitly
5. Use version control
6. Document adequately
7. Disseminate broadly
8. Conduct independent reviews
9. Test competing implementations
10. Conform to standards
Ten Simple Rules for Model Credibility

1. Define context clearly
2. Use appropriate data
3. Evaluate within context
4. List limitations explicitly
5. Use version control
6. Document adequately
7. Disseminate broadly
8. Conduct independent reviews
9. Test competing implementations
10. Conform to standards
MSM Task Forces

Task Force on Methodologies
- Cell-to-Macroscale Working Group
- High Performance Computing Working Group
- Multiscale Systems Biology Working Group
- Theoretical and Computational Methods
- Population Modeling Working Group

Task Force on Basic Science Applications
- Biomechanics Working Group
- Computational Neuroscience Working Group
- Integrated multiscale biomaterials experiment and modeling group (ImuBEAM)

Task Force on Dissemination
- Committee on Credible Practice of Modeling & Simulation in Healthcare Description
- Model and Data Sharing Working Group
- Public Dissemination and Education

Task Force on Clinical Translation
- Clinical and Translational Issues
- MSM for Medical Devices
- Multiscale Modeling and Viral Pandemics
Greater than the sum of its parts
Biological Spatiotemporal Scales

Hunter and Borg, Nature 2003

→ Individuals to Populations

Microscale

Mesoscale

 Macroscale
Piecing Together the Puzzle of Chronic Low Back Pain

A computer model may be able to inspire new insights and treatments.

The human back is a complex structure with bones, nerves, tendons, discs, and more — all places where something can go wrong and cause pain, which, for many people, becomes a long-term or chronic problem. Life stresses and other medical and mental health conditions aggravate the problem.

With so many pieces, it’s hard to get a holistic view of the puzzle or pinpoint the cause of the pain.

“People tend to focus on one aspect or another,” said Jeffrey Lotz, Ph.D., a medical engineer who studies back pain at the University of California, San Francisco. “Some people think it’s largely in the mind; some people think...”
Lower Back Pain
Built on the Foundation of the Biopsychosocial Concept
BACPAC Research Agenda – Goal 1
State-of-the-Art Model for Chronic Low Back Pain

• Develop a theoretical model for chronic low back pain

BACPAC Research Agenda – Goal 2
Identify factors that are predictive of treatment effectiveness for well-defined patient subpopulations

• Develop Testable Hypotheses

BACPAC Research Agenda – Goal 3
Develop an algorithm for multi-modal interventions for individuals with different phenotypes of chronic low back pain

• Design and conduct a large-scale adaptive cLBP trial that tests multiple bundled or sequential interventions
Future Robust Data Ecosystem

- All of Us Research Program
- 6 surveys
- Physical Measurements
- Structured EHR data
- GIS and Census
- Claims
- Bioassays from stored samples
- Notes, Labs, Imaging...
- Unstructured EHR Data
- Additional Digital Health Tech
- Additional Surveys
- Bioassays from stored samples
- Additional Measurements and Specimens
- Data Linkages
- -Omics
- All of Us RESEARCH PROGRAM

Current Data
**Inputs**
- Stimulus design
  - Frequency
  - Amplitude
  - Duration
  - Duty Cycle
- Electrode design and placement
- Animal model
- Disease model/phenotype
- Physiological recordings (for informing closed-loop stimulation)

**Outputs**
- Change in organ function (on- and off-target effects)
- Biomarkers
- Change in electrode interface (e.g. impedance)
- Plasticity/adaptation
CFDE

• Making data from Common Fund programs FAIR
<table>
<thead>
<tr>
<th>Experiment types</th>
<th>GTEx</th>
<th>HMP/IHMP</th>
<th>LINCS</th>
<th>Metabolomics</th>
<th>SPARC</th>
<th>4D Nucleome</th>
<th>MoTrPAC</th>
<th>Kids First</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole genome/exome sequencing</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Omni SNP Arrays</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16S metagenomic sequence</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epigenomics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcriptomics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA/miRNA/mRNA-Seq</td>
<td></td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>Array or L1000 platform based</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>Metabolomics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mass Spectrometry (GCMS, LCMS)</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipidomics</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteomics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mass Spec based</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>Microarray based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imaging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluorescence based</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology/Cytology</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radiology (MRI)</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In Vitro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binding Assays</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>Immuno Assays</td>
<td>✔️</td>
<td></td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In/ex vivo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural stimulation/neural activity recording</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
</tr>
<tr>
<td>High resolution manometry</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis Workflow</td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hardware Exists, Research Methodologies Lag

• Quantified self and health monitoring technologies are evolving rapidly; data is often proprietary

• Challenges
  • Methods for dealing with data size and complexity
  • Validation of models for application in clinical research and trials
  • Multidimensional interpretation of data
  • Ethics of use, provenance of data, and regulatory strategies
Biomedical and Behavioral Research Challenges

Data Sources
- Heterogeneous data sources
- Unstructured data sources
- Lack of data on social determinants of health
- Measurement issues (e.g. incompleteness, inaccuracy, imprecision in self-reported data)
- Privacy and security
- Cost
- Limited adoption of common data models

Study Designs
- Semantic data integration (i.e. linking data elements by their meaning)
- Large longitudinal cohorts
- Ontology integration
- Ontology appropriateness (e.g. ontologies made for billing vs. for diagnostic purposes)
- Semantic interoperability- common language
- Automated study design

Prediction Modeling
- Biases of all sorts (e.g. protopathic)
- Confounding
- Causal inference vs. mechanistic
- Black-boxes vs. white-boxes (i.e. interpretability vs. performance)
- Complexity-based model selection
- Benchmark development
- Pragmatic interoperability (reproducibility, replicability, generalizability)- different labs able to get the same result

Translational Relevance
- Limited individual empowerment
- Disconnect from relevant clinical research
- Personal health record/health avatar (besides provider’s electronic records)
- Acceptance of artificial intelligence as integral part of doctors’ tools
- Learning systems
- Ethical usage and dissemination of modelling algorithms
- Redefining disease phenotype

from Prosperi et al. BMC Medical Informatics and Decision Making (2018)
HOW DO WE GET THERE?
The NIH Artificial Intelligence Initiative through 2028
(acronym TBD)

To Propel Progress in Biomedical Research through NEXT-GENERATION AI
A follow-up from the July 2018 AI Workshop,
https://datascience.nih.gov/community/2018biomedAI

https://videocast.nih.gov/watch=35426
(start at 1 hour, 2 minutes)
Parallel Revolutions: Fusing Biomedicine and Machine Learning (12/13/2019 ACD Report)

**Data Generation**
more data about the biology and health of more individuals than ever before

**Data Analysis**
machine learning, other forms of artificial intelligence, cloud computing

**ML-BioMed**
- biomedical experiments*
  that are designed for ML
- ML that's designed for biomedical experiments*

*Note: Biomedical experiments include **biological** and **behavioral** studies
Support flagship efforts that generate large-scale experimental data, with billions of data points designed to:

i. be well-suited for ML analysis and inference
ii. address key biomedical challenges
iii. stimulate new approaches in machine learning

And that implement processes designed to:

i. develop improved criteria and technical mechanisms for data access
ii. strengthen ethical criteria for dataset use (consent, privacy, accountability, ...)

Projects should:

- address key biomedical challenges using ML methods
- advance ML methods for future use in biomedicine
- produce transformative data sets, designed with ML in mind
- propel new ways to gather massive data in biomedicine
- involve strong engagement from leading ML researchers

Project review should:

- incorporate expertise in ML as well as traditional biomedical domains
Overall Initiative Goals

- Establish a launchpad for widespread adoption of Next-Generation AI
- Create next generation AI-driven scientific design and assessment frameworks

- Enable transformative data collection around grand challenges in biomedical research
  - Challenges that are currently beyond our human intuition and require next-generation AI approaches to solve
  - All future challenges to use adoptive framework
Artificial Intelligence – what’s next-gen?

- Self-driving cars
- Facial recognition tools to predict depression and mental health
- Detection of cancerous pulmonary nodules on chest X-rays
- Improved accuracy and speed of interpretation of pathology slides
- Classification of skin cancer by image analysis
- Diagnose heart attack from ECG data
- Finding diminutive (<5mm) polyps in a colonoscopy
- Diagnosing eye conditions from retinal fundus photographs

1) Remove the black box - explainable AI
2) Remove fragile AI – impenetrable security
3) Remove bias – ethics in AI
4) Remove big data – learn from small data
5) Address physics of AI - infrastructure
Narrow AI

One Study (data source) – multiple data types

One Study (data source) – one data type

One Study (data source) – one data type

Current AI approaches

Broad AI

One Study (data source) – one data type

One Study (data source) – one data type

One Study (data source) – one data type

Future AI approaches
Who/What needs to come together?

- Biologists/Biomedical Scientists
- Engineers
- Mathematicians
- Statisticians
- Computer Scientists
- Physicists
- Chemists
- Clinicians
- Economists
- Philosophers
- Anthropologists

Next-generation AI methods?

- Transfer learning
- Artificial general intelligence
- End-to-end learning (in DL)
- Tabula rasa learning theory
- Bayesian networks and inference
- Pearl-esque probabilistic causal learning
- Monte Carlo simulation and tree search
- Hypothesis-free, unsupervised DL
- High-scale modeling for prediction and forward simulation
- Quantum-inspired optimizations, including sampling, minimization, and training neural networks

Prosperi et al. BMC Medical Informatics and Decision Making (2018)
Grand Challenges that integrate all types of biomedical and behavioral data to predict health outcomes

From: Big data hurdles in precision medicine and precision public health
Prosperi et al. BMC Medical Informatics and Decision Making (2018)
Some Environmental Factors Influencing Outcomes

Precision public health. Community, societal and ecological factors must be accounted on top of the individual-based, fine-grained approach for precision medicine. The map is an edited version of a Wikimedia Commons image (https://commons.wikimedia.org/wiki/File:United_States_Administrative_Divisions_Blank.png, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported)
Grand Challenge Use Cases

- Connecting multiple pieces of information for research
  - Integrating heterogenous, variable, uncertain data
  - Finding the hidden signals to derive knowledge and insight

- Wearables give a “movie for health”
  - How to read between the lines of data
  - Finding the signal we never saw before

- Generalizing from known diseases to rare diseases
- Finding the mechanisms of health restoration
Discovering Theories & Knowledge → INSIGHT

Model Integration
- Theoretical Constructs
- Model Parameters
- Model Uncertainty

Data Integration
- Competing Theories
- Sparse Data
- Missing Data

Data Imputation
- Uncertain Data

Visualization
- Sparse Data
- Missing Data

Models & Data across species, spatiotemporal scales, behavioral tasks

Reveal Emergent Dynamics, Hidden Rules → Integrated Theories

Development and implementation of a machine-driven knowledge integration process for discovery
1960: A Vision for Cognitive Assistance

“In not too many years, human brains and computing machines will be coupled together very tightly, and the resulting partnership will think as no human brain has ever thought!”

--JCR Licklider
Visionary psychologist and computer scientist
Funded research that led to most of modern computing

Tesla Truck, November 2019

Google Car, 2018

Oakley Radar Pace

• https://www.youtube.com/watch?v=-S4V1TS4yFk

Radar Pace, available online
AI can change how we gather data, not just how we interpret it.

Learned reconstruction with neural networks

DK Sodickson, P41 EB017183-05S1
Direct to information

DK Sodickson, P41 EB017183-05S1
AI to Extract Human Intelligence (not human stupidity)

- Integrating the right types of data
  - at the right time
  - within the right context at (this point in time)
- Avoiding human biases
  - coding for ethics and (human?) learning
  - AI research for ethics
- Incorporating the totality of the data
  - without collecting ALL the data
  - Stitching together SMALL data
- Re-engineering the Future of Health
  - Understanding the mechanisms of prevention, diagnosis and treatment
  - Lowering healthcare costs
Vision: To Propel Progress in Biomedical Research through NEXT-GENERATION AI (beyond Narrow AI to Broad AI)

Culture Change: Biomedical experiments designed for next-gen AI → AI designed for biomedical experiments*

3 Pillars: People (multiple disciplines), Data (transformational), Ethics (data bias and transparency)

Goals/Outcomes after 7 years (FY21-27, ~$200M):
- AI Design Centers for the Biomedical Community — data design and assessment, ethics and training
- New “Gold Data” that can be mined with future AI methods
- Ability to “stitch” Gold Data with existing data (across sites, protocols, processing methods)
- Next generation discoveries for biomedical research, powered by next-gen AI

Partnerships: DARPA, NSF, DOE, FDA, ...

* Includes biological and behavioral studies

Immediate Timeline:

**October 27-28, 2020:** Community Workshop in partnership with DARPA Synergistic Discovery and Design (SD2) program

**Fall 2020:** Release of Funding Opportunities for DEFINE AI Design Centers inspired by Biomedical Grand Challenges

**Fall-Winter 2021:** Formation of Multidisciplinary Teams, Vet Grand Challenge Ideas → Online breakout groups for each Grand Challenge idea
Thank You!

Grace C.Y. Peng

grace.peng@nih.gov