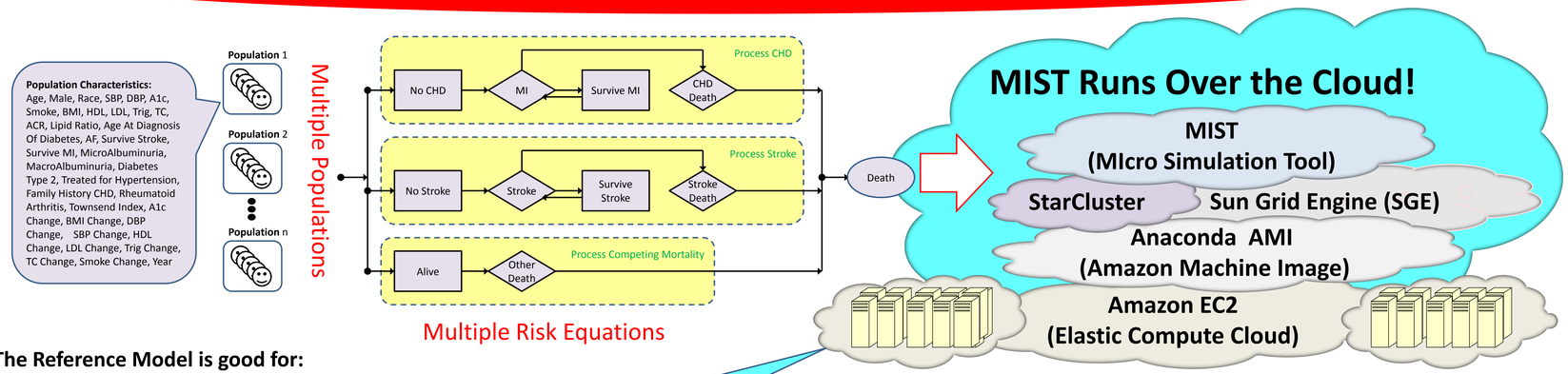
The Reference Model for Disease Progression Sensitivity to Bio-Marker Correlation in Base Population
The Reference Model Runs with MIST Over the Cloud!

Jacob Barhak Ph.D.

http://sites.google.com/site/jacobbarhak/

IMAG Multiscale Modeling (MSM) Consortium Meeting 2013 - 2-3 October

In a Nutshell: A League of Disease Models using High Performance Computing



The Reference Model is good for:

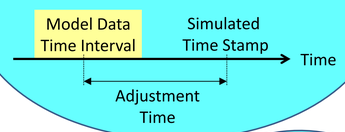
- Economic forecast: Cost/Quality of Life
- Comparison amongst studies
- Evaluation of models – "consumers report"
- Better model development
- Visualizing the bigger picture of our understanding of disease progression
- MultiScale Modeling: individual to population to multi-population

The Reference Model Accomplishments so far:

1. Disease models go out of date and need for temporal correction
2. Identify models that behave better/worse
3. Using end of trial biomarkers in equations is beneficial
4. Identify which diabetic studies are easier / harder to explain
5. Bio-marker correlation is not the main reason for our gaps in understanding disease progression

Reproducible!!!

1. Temporal Correction for Model Outdating / Treatment Improvement:



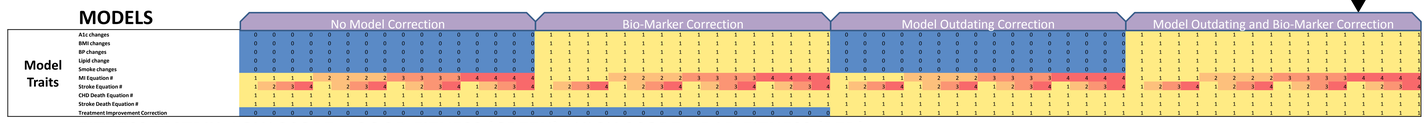
New!!!

Possible Future Directions:

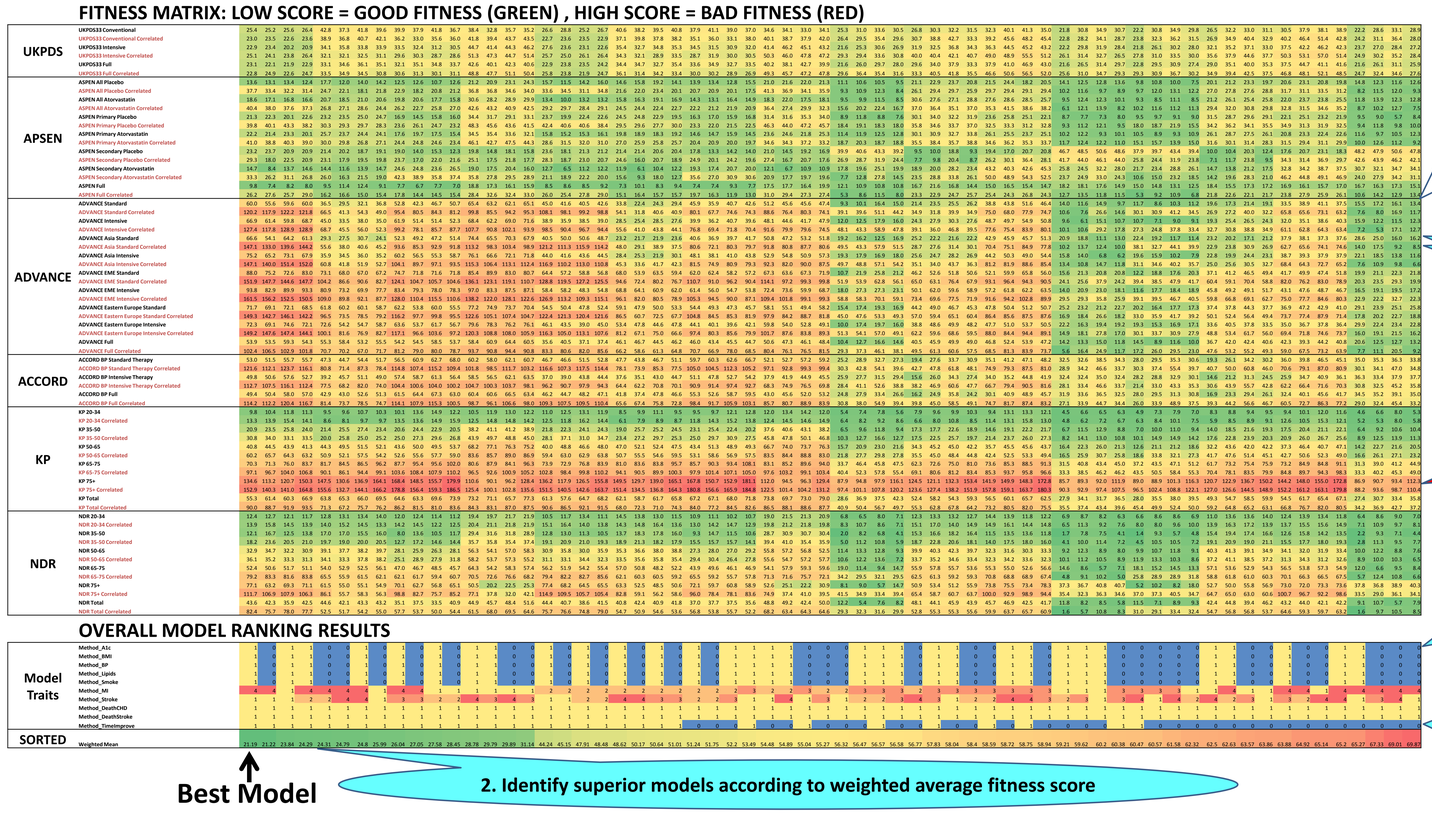
- Optimizing MIST performance
- Adding more populations
- Adding more risk equations
- Adding more disease processes
- Additional analysis modes
- Using Individual data
- Formulating alternative fitness functions

Best Model

MODELS



FITNESS MATRIX: LOW SCORE = GOOD FITNESS (GREEN) , HIGH SCORE = BAD FITNESS (RED)



Best Model

2. Identify superior models according to weighted average fitness score

OVERALL MODEL RANKING RESULTS

1. The best 24 model variations include the temporal correction for model outdating/ treatment

2. Using end of trial biomarkers produces a slightly better model

3. Fitness Diff due to correlation: Avg. 15 ± 17SD 1000 individuals

4. No model used can explain the phenomenon observed in this cohort

5. Fitness Score:

In Lay terms the fitness score shows the difference between simulation results and end of trial results for 1000 patient s

Better fitness

Worst fitness

The Reference Models for disease progression [1-3] uses High Performance Computing (HPC) to simulate a league of published disease models to compete for fitness to known clinical trial results. The resulting color coded fitness score matrix serves the modeler as a reference to improve model design and improve understanding of different models. The Reference Model regenerates baseline population data from publicly published data and does not require access to individual population information. Therefore the base for information is large, combining information from multiple trials, which provides a wider global view of phenomena observed. Using a Multi Scale modeling approach simulations are executed at the individual scale to reproduce information at the population scale and allows drawing conclusions at the multi-population scale.

Never the less, the use of aggregate data instead of individual data brings up questions regarding the sensitivity of the model to assumptions and unpublished information such as bio-marker correlations. Fortunately it is possible to address these questions using HPC by running the assumptions and their fitness to observed results.

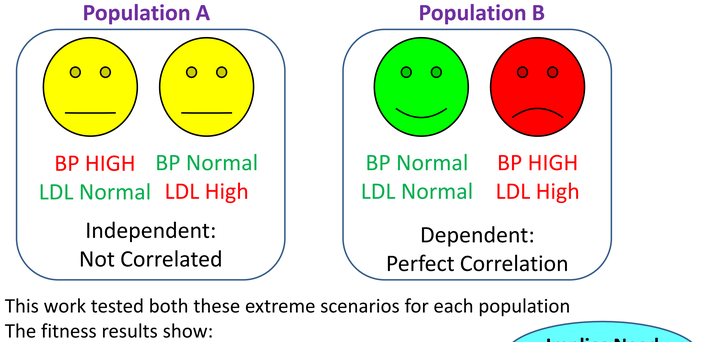
This work demonstrates this by testing two extreme scenarios of biomarker correlation within populations: 1) Independent Bio-Markers with No Correlation, 2) Dependent Bio-Markers with Perfect Correlation. 2 X 34 cohorts of 6 diabetic studies :UKPDS, ASPEN, ADVANCE, ACCORD, KP, NDR, were generated from distributions using those scenarios and tested against 64 model variations composed of cardiovascular risk equations and modeling assumptions [1,2]. The results show the behavior of the model variations within the assumption scope and result rankings help identify superior models.

Results consist of 4352=2x34x64 scenario simulations of 1000 virtual individuals and 10 Monte Carlo repetitions each for 10 years ~ 0.4 Billion virtual patient years. The use of HPC allowed running these computations in reasonable time. This simulation took about 36 hours on the cloud by leasing a computer cluster composed of 160 cores using 20 machines with 8 cores each.

The Reference Model uses free Python based framework of software tools to run these simulations. MIST (MIcro Simulation Tool) [4-5] is used to run simulations in HPC environment. MIST runs over the cloud with the help of StarCluster [6] and an Anaconda AMI (Amazon Virtual Machine) [7].

Lay Explanation

- Clinical trials do not typically publish correlation amongst biomarkers
- Therefore it is impossible to distinguish between the following populations:



- This work tested both these extreme scenarios for each population
- The fitness results show:
  - Correlation matters since a range in results is visible
  - Correlation does not change the big picture much
- Conclusions:
  - **Modeling at the individual level is important and makes a difference**
  - **Correlation is not the main hurdle in understanding disease progression**

Implies Need for HPC

Better models needed

References:

[1] J. Barhak, The Reference Model Scores Fitness of Models and Populations. Poster Presentation. ISPOR 18th Annual International Meeting, May 18-22, 2013, Sheraton New Orleans, New Orleans, LA, USA.
Poster: http://sites.google.com/site/jacobbarhak/home/PosterTheReferenceModel_ISPOR_Submit_2013_05_14.pdf

[2] J. Barhak, The Reference Model: Improvement in Treatment Through Time in Diabetic Populations, The Fourth International Conference in Computational Surgery and Dual Training. The Joseph B. Martin Conference Center at Harvard Medical School. Boston, MA, USA. December 9-10-11, 2012.
Video: http://web.cs.uh.edu/~cosine/?q=node/140 ,
Presentation: http://www2.cs.uh.edu/~cosine/talks_cosine4/monday/MultidisciplinaryTalks/2_JacobBarhak.pptx

[3] J. Barhak, The Reference Model for Chronic Disease Progression. 2012 Multiscale Modeling (MSM) Consortium Meeting, October 22-23, 2012,
Abstract: http://www.imagwiki.nibib.nih.gov/mediawiki/images/7/77/Reference_Model_for_Chronic_Disease_Progression_Barhak.pdf
Poster: http://www.imagwiki.nibib.nih.gov/mediawiki/images/c/c4/PosterTheReferenceModel_IMAGE_MSM_Submit_2012_10_17.pdf

[4] J. Barhak, MIST: Micro-Simulation Tool to Support Disease Modeling. SciPy, 2013, Bioinformatics track,
Abstract: https://github.com/scipy/scipy2013_talks/tree/master/talks/jacob_barhak
Video: http://www.youtube.com/watch?v=AD896WakR94

[5] MIST - Micro Simulation Tool software in the GitHub repository. Online (Accessed 12-Aug-2013): https://github.com/Jacob-Barhak/MIST

[6] StarCluster home page. Online (Accessed 12-Aug-2013): http://star.mit.edu/cluster/

[7] B. Zaitlen, StarCluster Anaconda. Online (Accessed 12-Aug-2013): http://continuum.io/blog/starcluster-anaconda

Acknowledgments:

The GPL modeling framework was supported by the Biostatistics and Economic Modeling Core of the MDRTC (P60DK020572) and by the Methods and Measurement Core of the MCDTR (P30DK092926), both funded by the National Institute of Diabetes and Digestive and Kidney Diseases. The modeling framework was initially defined as GPL and was funded by Chronic Disease Modeling for Clinical Research Innovations grant (R21DK075077) from the same institute.

The Reference Model was developed independently without financial support.