The Reference Model for Disease Progression Handles Human Interpretation

Jacob Barhak
Jacob Barhak
Austin
jacob.barhak@gmail.com

ABSTRACT

The Reference Model for disease progression is an ensemble of models of cardiovascular disease and diabetes. To date, it is the most validated cardiovascular diabetes model known. It accumulates information from multiple sources including publications and ClinicalTrials.Gov. The model can visually show how well our cumulative knowledge can explain phenomena observed in clinical trials. Our current computational understanding has shown a gap of comprehension. Explanations to this gap exist in some cases. However, those use human explanation using human language while a computational model requires hard coded rules. This gap between human and machine comprehension needed a remedy.

The Reference Model was therefore recently equipped with the ability of include human expert understanding. Moreover, the model can now accept interpretations from multiple human experts and integrate it with its accumulated knowledge and its optimization process.

This new capability will be discussed in the paper and results will be presented.

ABOUT THE AUTHORS

Jacob Barhak is a Computational Disease Modeler focusing on computer comprehension of clinical data. He has diverse international background in engineering, computing science, and disease modeling. He was educated at the Technion Israel Institute of Technology and worked at the University of Michigan. The Reference Model for disease progression was self developed in 2012 and he currently pursues this development effort as an independent researcher. His efforts include standardizing clinical data as seen in ClinicalUnitMapping.com. He is the developer of the Micro Simulation Tool (MIST). His work was internationally exposed in multiple venues. His disease modeling work was repeatedly presented in multiple venues, including MODSIM, SummerSim, I/ITSEC, MSM/IMAG meeting, PyData & Pycon Israel. Dr. Barhak is an advocate of non-blind public scientific review processes and active in organization work, including SummerSim and the Population Modeling working group. He is active within the python community and runs the Austin Evening of Python Coding meetup. His activities and publications are summarized in the following web page: http://sites.google.com/site/jacobbarhak/

The Reference Model for Disease Progression Handles Human Interpretation

Jacob Barhak
Jacob Barhak
Austin
jacob.barhak@gmail.com

INTRODUCTION

Computational Disease Modeling is a field where computational models attempt to predict outcomes for a population or an individual by using computer models. Those models many times are expressed as risk equations that attempt to predict the probability of an outcome in a patient with specific characteristics e.g. (Stevens, 2001) , (Wilson et. al., 1998). For example what is the probability of a patient experiencing stroke in 10 years given their age, blood pressure and other parameters. Those risk equations are typically developed by a modeling group that has access to longitudinal data of patient data.

Typically patient data in the medical world is highly restricted and is rarely shared with other groups, so publishing the risk equation/model is one way of sharing knowledge that does not compromise the restricted data. However, combining this knowledge was very limited for many years. Assembly attempts by some groups included assembling their own equations to models that predict multiple outcomes (Clarke et. al., 2004), (Hayes et. al., 2013) and others assembled equations from multiple sources into one model (Barhak et. al., 2010). Yet at this earlier time, global assembly of information was not possible.

A lot of progress was done in the diabetes modeling community and modelers started comparing their model in the Mount Hood challenge (Mount Hood 4 Modeling group, 2007) where multiple modeling groups would meet to compare and contrast their models. However, the models constructed by multiple teams were different and results varied across multiple groups when validation challenges were attempted. In validation challenges, baseline population statistics were given and modeling teams were competing in how close they can predict the outcomes for that populations. Populations typically represented clinical trials with a few executions, so summary data was publicly available. Despite the availability of data, the predictions provided by multiple teams varied and were not accurate. Moreover, each time a modeling challenge was introduced, there was no continuity to previous challenges and validation against populations from previous challenges was not required in a newer challenge.

Although attempts were made to standardize input data for challenges, the process was a human intensive process focused on the modeling teams making assumptions and interpreting ambiguous data rather than an organized procedural process that can be automated.

The inability of the diabetes modeling groups to replicate known outcomes and the variety of models inspired the author to take a new approach that will merge information from multiple sources and validate them against multiple sources in an automated manner. The Reference model was the solution.

THE REFERENCE MODEL FOR DISEASE PROGRESSION

The Reference Model started with the idea to automate the Mount Hood challenge. Instead of multiple groups of humans meeting once every other year and preparing for a few months for one challenge, a machine can receive all models and run them on a the same standardized inputs. This can happen continuously and also allow accumulation of knowledge in one place so that multiple challenges can be stored together. Yet once the problem was formulated for a computer, it opened many more possibilities for accumulating knowledge as will be described later. Yet we are ahead of ourselves and should start with the first model version.

The Reference Model was created in 2012 as an automated mini replica of the Mount Hood Challenge aimed at diabetic populations. The model included 3 processes coronary heart disease, stroke, and competing mortality. This structure of the model was relatively simple as shown in Figure 1. The arrows in the model diagram represent transitions between disease states. During simulation a random number is picked for each active state and it is compared to the risk equation that represents a threshold for transition. This way the model decides if an individual moves to a different state or stays in the same state for that time step. This is repeated for each individual in the population. At the end of simulation the model outcomes are compared to known population outcomes to figure out how good the model is, we will call this number fitness.

Despite its simplicity, the model allowed complexity that was not possible with the human based challenges, it allowed assembling a model using different risk equations. Each transition probability could be represented by more than one risk equation. The Reference Model was therefore not one single model, it was an ensemble model that is composed of many models. However, initially the full potential of the model was not realized since the different models were made to compete - very similar to what was done at the Mount Hood Diabetes challenge. Each time a simulation executed, a different equation was chosen for each transition probability. For example Equation A would be chosen for the probability for Myocardial Infarction (MI) and Equation E was chosen as the probability of Stroke - denoted by the combined model AE. We could contract multiple such models: AE, AF, AG, AH, BE, Bf, BG, BH, CE, CF, CG, CH, DE, DF, DG, DH and this number would grow up exponentially and therefore High Performance Computing (HPC) was required to run all those models and figure out which one represents best the phenomena observed in the population. And this was executed for multiple populations to figure out the model that behaves best for all populations. This approach was competitive and although it allowed accumulating more knowledge than the human challenges that lacked consistency by removing previous challenges, it did not reach full modeling potential.

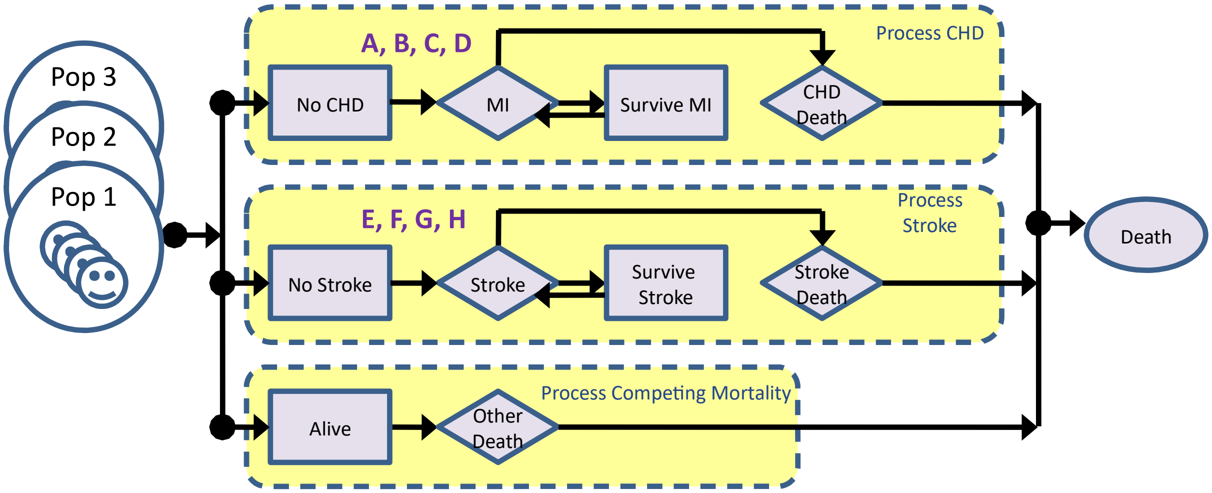


Figure 1. The Reference Model Diagram

The full potential was realized after the number of models and populations grew, it was then necessary to switch to a much better approach that utilized the full potential of the ensemble model - a cooperative approach. The key observation was that no one model is perfect and all models should be treated as assumptions rather than absolute truths and we wish to merge assumptions together so those will cooperate. In this cooperative approach, all risk equations contributed to a combined risk according to their influence. For example, for the MI probability equations was assigned a weight and the combined probability for a transition was w1A+w2B+w3C+w4D where the coefficients w1,w2,w3,w4 are scalar weights that represent the influence of a certain equation. The Reference Model then represented an infinite number of models that represent disease progression based on risk equations as basis functions. The modeling space then became a continuous function that can be optimized using mathematical optimization techniques that are very similar to those used in training neural networks (Barhak, 2016). The solver was named as: "assumption engine" since it figures out which assumptions work better together considering the data and query. This cooperative approach allowed creating models that behave better than any of the original risk equations alone. Moreover, it could be used in combination with the competitive approach for testing assumptions that are not continuous in nature.

Information accumulation went beyond multiple models being integrated into one ensemble model. Much important information is provided by population data that was also incorporated. The Reference Model started with validating against a few past populations from the Mount Hood challenge and the literature. This number increased with additional challenges. Yet unlike the human challenges that did not retain memory from previous validations, the ensemble model retained those and this data was accumulated rather than forgotten. The Reference Model uses population data that was publicly composed of summary statistics rather than restricted individual data that is typically not released. The model needed to simulate populations that matched the demographic of those population cohorts. This was done by sophisticated population generation driven by the MIcro Simulation Tool (MIST) (Barhak, 2013) that served as the computational engine behind the model. Since population generation was a Monte Carlo random process, there was a need to improve accuracy to better match population statistics. This was accomplished using Evolutionary Computation algorithms (Barhak & Garrett, 2014). However, when the model grew, the amount of code that was required became unreasonable and object oriented population generation code was introduced to allow efficient and compact population generation (Barhak, 2015).

Yet even with efficient ways of recreating populations, the process was slow - it took roughly a week of work to recreate one population from a publication and much of this work relied on copying numbers from published papers and writing generation code. This was remedied when an interface was created for ClinicalTrials.Gov that reduced the time required to add a population to a few hours per population, while eliminating human error.

ClinicalTrials.Gov is the registry where clinical trials report their structure and results. This database growth is driven by U.S. law and already holds over 300,000 clinical trials with over 41K clinical trials with results. Results data that was previously published without uniform format in scientific journals is now entered into a database. An interface was created that allows the modeler to use extracted data and semi-automatically create populations that can be simulated by the ensemble model. This interface caused a dramatic increase in the amount of knowledge held by the model. The Reference Model then became the most validated diabetes cardiovascular disease (CVD) model known worldwide, bypassing the previous champion - the Archimedes model (Eddy & Schlessinger, 2003). Today, there is no other known CVD diabetes model that accumulates information from so many sources with validation.

With so much information, it was then possible to visualize our computational knowledge gap. This gap shows how the most fitting model assembled from the base equations fits all clinical trials. This was presented using interactive techniques based on Python visualization libraries (Bokeh, Online), (HoloViz, Online).

With so much information assembled, it was possible to analyze data in ways not possible before. For example the rate of improvement of treatment in CVD diabetic death could be assessed, so a similar idea for Moore's law could be defined. the model discovered that diabetic CVD death probability decreased roughly by half every 5 years as calculated using 3 decades of models and populations (Barhak, 2017). Life tables were published using two scenarios: 1) using improvement rate into account, 2) not correcting for treatment improvement rate. This was just one example of what is possible when information from multiple sources is centralized in one ensemble model.

However, despite all the progress made, information arriving from multiple sources is still prone to human error despite capabilities of detecting wrong equations. Even strict testing was shown to bypass a few errors each year. For example, the results in this paper correct a row shift and a mismatch in a result matrix that was introduces by human errors in the two last published versions However, more automation and accumulation of knowledge will eventually diminish a possible error to be negligible and hence the need to go away from human focused modeling to automated modeling. For example, the erroneous outcome entry in the last publication (Barhak, 2020) is only one from 120 outcomes entries and therefore if its influence is not strong when comparing results and can be considered negligible. Moreover, one equation know to be erroneous is rejected by the model on the first iteration, thus demonstrating how accumulated knowledge effectively reduces error.

However, even if the process becomes highly automated, humans still need to be involved in the modeling process. Humans, just like models, have different opinions and many times there is no easy way to measure the accuracy of those opinions. Since humans need to drive the modeling process, instead of the human being concerned with performing repetitive tasks, humans should be focused on looking at data and results. In this paper we introduce one way of doing this by including human interpretation to deal with ambiguous or fuzzy data while employing machine learning to figure out the best fitness when considering interpretation by a team of experts.

HANDELING HUMAN INTERPRETATION

When transforming medical data into a model there are many human considerations taken. Many of those are not computational in nature and relate more to understanding texts. Despite advances in Natural Language Processing (NLP) machines still cannot perform human language interpretation properly and computational model creation based on such data is even a harder task. However, for a computational model that validates predictions to outcomes, it is possible to pose the problem in a way a machine can comprehend.

Outcomes of a clinical trial are typically counts of a certain observed phenomenon, for example a stroke. However, a stroke can be defined in many ways and therefore different trials may report the same outcome differently. Some times the definition of an outcome is made using International Statistical Classification of Diseases (ICD) codes. However, even when well defined in one ICD version, the definition may change in another ICD version. For example in (Clarke et. al. 2004) ICD 9 Stroke is defined by as (ICD-9 codes ≥430–≤434.9, or 436) . However, when translating to ICD 10 codes, the list closely translated to I60.9, I61.9, I62.1, I62.00, I62.9, I65.1, I63.22, I65.29, I63.139, I63.239, I65.09, I63.019, I63.119, I63.219, I66.09, I66.19, I66.29, I63.30, I66.9, I63.40, I66.9, I67.89. Only looking at the first code of ICD9-430 the definition is "Subarachnoid hemorrhage" while the ICD 10 I60.9 equivalent is defined as: "Nontraumatic subarachnoid hemorrhage, unspecified" these small changes in definition eventually cause confusion for a machine when the word stroke appears in a published report. Although a human will be able to explain what a stroke means, for a computer a different definition of the words that describe stroke or a different code list will be hard to decipher.

This problem aggravates further since in tables that describe clinical trial results, the ICD codes that define a specific outcome are not specified directly and although many times those can be found after an exhaustive human search in the trial protocol or in another location in a related publication, many times there are differences in reporting outcomes between trials. The problem aggravates even further in composite outcomes such as cardiovascular disease (CVD) that include many other outcomes including MI and stroke. The definitions of outcomes sometimes even differs within the same clinical trial that reports the same outcome using different definitions. For example the RECORD clinical study (ClinicalTrials.gov - NCT00379769, Online) reports the same outcome twice using two different criteria: 1) "Independent Re-adjudication (IR) Outcome: Number of Participants With a First Occurrence of a Major Adverse Cardiovascular Event (MACE) Defined as CV (or Unknown) Death, Non-fatal MI, and Non-fatal Stroke Based on Original RECORD Endpoint Definitions" 2) "Independent Re-adjudication Outcome: Number of Participants With a First Occurrence of a Major Adverse Cardiovascular Event (MACE) Defined as CV (or Unknown) Death, Non-fatal MI, and Non-fatal Stroke Based on Contemporary Endpoint Definitions". Although this trial has properly reported the outcomes using multiple interpretations, it is unclear how to compare those outcomes to a different trial and how to validate those against simulated model outcomes, especially when an ensemble model is considered - the description is not traceable back to quantifiable definitions and therefore hard to a machine.

Similar definition changes are not uncommon, the definitions in medicine change constantly even outside cardiovascular disease. For example the definition of sepsis was changed numerous times in a few decades as seen in (Gary et. al., 2016), (Wentowski et. al., 2018). And since the model accumulated clinical information spanning over several decades, there is a necessity to add human interpretation to outcomes being used for validation.

However, note that humans may not always understand the data the same way, and human interpretation of the same outcome may differ from one expert to another. The example of the RECORD study(ClinicalTrials.gov - NCT00379769, Online) discussed earlier shows how the same outcomes are interpreted differently and numbers differ. So we wish to be able to add human interpretation of outcomes from multiple experts that will evaluate possible ambiguous information.

In the past, the Delphi method (Wikipedia - Delphi, Online) was used to assemble information from multiple experts. One example of a derivative of the method was used for mental health modeling (Leff et. al., 2009). However, those techniques are human based and require human feedback and reiteration which is time consuming. We want a technique that takes human inputs and allows merging it efficiently with the power of machines to compute and validates the assumptions that experts make.

MATHEMATICALY HANDELING HUMAN INTERPRETATION

Human interpretation can potentially be added to any aspect of modeling, yet it was initially applied only to outcome interpretation. Consider the following notations:

R - simulation result - this is the number the model generates after Monte Carlo simulation.
T - expected target outcomes - these are the numbers that appeared at the clinical trial results - our ground truth
$H_i(T)$ - Human interpretation of T by expert i - representing what the expert thinks the ground truth should be
D - difference between ground truth and simulated results - this is the fitness/error we wish to me minimal.
w_i - the weight we assign to expert i interpretation - it represents how much we believe that expert

The basic idea is to find the best balance of experts that will increase the prediction accuracy of the simulation. The Reference Model uses a fitness engine that calculates the difference between simulated results and expected outcomes and attempts to optimize it. Without Human interpretation, this would be defined as:
$D = T\text{-}R \ \rightarrow min$
However, when we introduce human interpretation, this difference becomes a weighted sum considering all experts:
$D = \Sigma\, w_iH_i(T) - R \ \rightarrow min$
subject to:
$\Sigma\, w_i = 1$
$w_i \geq =0$

The constraints make sure that the combined weighted interpretation of all experts is within the convex hull of all the interpretations given and that no interpretation given by an expert is considered as false - at worse case the interpretation is incorrect if $w_i=0$. In simpler words it means that the minimum and maximum after accounting for all expert interpretations will be bound by the largest and smallest outcome interpretation of the experts.

Also note that the assumption engine already includes a very similar formulation where w_i also decides the level of influence for a certain model equation as described before when assembling the ensemble model: $w_1A+w_2B+w_3C+w_4D$. In fact the interpretation of the expert can be considered part of the modeling assumptions that require optimization. The only difference is that to calculate the fitness D for interpretations there is no need to recalculate the results R - which involves the entire simulation that involves validation of the population against the model - which is time consuming and typically takes about 16 hours on a 64 core machine to account for all variations and populations. Instead, we can quickly calculate all variations of interpretations very quickly without the need to recalculate R. And since the assumption engine already uses gradient descent optimization to improve w_i for model components (Barhak, 2016), we just add an extension of w_i related to human interpretations to the solution vector and use the same solver rather than decoupling the human interpretation handling from the model assumptions handling. Here is proof that this decoupling is possible.

Lets call the Difference between ground truth and human interpretation of expert i as :
$D_i = \ w_i(H_i(T) - R)$
We will defined the combined difference instead as:
$D = \Sigma D_i \ = \Sigma w_i(H_i(T) - R)) = \Sigma(w_iH_i(T) - w_iR) = \ \Sigma(w_iH_i(T)) - \ \Sigma(w_iR) = \ \Sigma(w_iH_i(T)) - R{*}\Sigma(w_i)$
Since $\Sigma(wi) = 1$ we get again: $D = \Sigma(w_iH_i(T)) - R$, which means that we can decouple the simulation from interpretation for the sake of determining interpretation weights of experts for optimization purposes. So when running the code we use the $D = \Sigma D_i$ formulation to deduce the combined interpretation difference.

Yet this description is still somewhat simplified compared to actual code that implements the simulations since each outcome appears in some populations. The actual way that experts interpret outcomes is by looking at the outcome description of a specific trial and expert i assigns a scalar number z_{ij} associated with outcome for a specific trial j. this number is used to adjust the ground truth T_j for all cohorts of trial j so that $H_i(T_j) = z_{ij} {*}T_j$. If $z_{ij} = 1$ it means that the expert believes that the reported outcomes match the model definition of the same outcome. If $z_{ij}<1$ it means that the outcome defined by the study over-counts incidence compared to how the model views the definitions. if $z_{ij}>1$ then the study results in the publication does not include some outcomes defined by the model and the under-counted observed outcome should be increased to match the model definition. Also note that the model definition includes multiple merged models with different weights. Since all weights are optimized, the most fitting balance of all interpretations and assumptions is created - optimally mixing the model and expert definitions.

IMPLEMENTATION

The Reference Model code was modified to incorporate human interpretation optimization as described before. As explained earlier, the code change could be merged with existing optimization code. Therefore, a lot of effort was put into handling the data. However implementation included multiple other changes. One minor change added warning code to isolate an issue with an equation that was previously marked as wrong by the assumption engine.

The major change was that all outcomes that were reported by all studies entered into the system were revisited. Those study outcomes were previously matched with model definitions of outcomes using free text that explains the modeling assumption and as a table matching the outcome to ICD codes, this was done for MI, Stroke, CVD and mortality and their combinations. Much effort was put previously in documenting the modeling assumptions regarding outcome definitions, yet this was only a documentation file. In the new version this documentation was adapted to a matrix of human assigned values that can be incorporated into computation. Each row in the matrix of values contained a single outcome extracted from a certain study including human explanation. There were many columns in that matrix, most of which contained documentation. A few numeric matrix columns were added to contain numeric human interpretations. Ideally each column should have represented a different expert opinion on how well the study outcome matches the model definition as a positive number around 1. Those values correspond to the z_{ij} values that go into computation.

In this publication, only the author wrote all interpretations while trying to imitate 6 experts with different opinions both conservative and liberal - we mark them as 1-6 in Table 1 below, each time making other assumptions trying to simulate conservative experts that stick to the textual definitions and emphasize the difference by assigning numbers farther than 1 in a direction that fits their "assumed personality". More liberal experts may accept differences in text more easily and report numbers closer to 1. Note however, that death was considered absolute outcome that all experts gave the interpretation of 1. The first interpretation in the interpretations columns was full of 1 values indicating that model outcome matches study outcome. Note that Table 1 provides only a small glimpse into the interpretations used for a small number of the 120 outcomes used in the simulation - just to illustrate the procedure.

Table 1. Small subset of the interpretation data

Study	Outcome	Expert Interpretations						Reference	Comment
		1	2	3	4	5	6		
UKPDS33	Death	1	1	1	1	1	1	(UKPDS,1998)	All deaths counted
ADDITION	MI	1	1	1.2	0.8	1.2	0.8	(Griffin et. al., 2011)	Exact detailed definition is not available in the paper, and since it is a multi national trial, it is assumed that there is some variability beyond MI+Stroke
ADDITION	Death	1	1	1	1	1	1	(Griffin et. al., 2011)	Death is absolute
RECORD	MI	1	1	1	1	1.05	0.95	(ClinicalTrials.gov NCT00379769 , Online)	Word description is very specific and short with little room for interpretation of MI
THRIVE	CVD	1	0.8	1	0.6	1	0.6	(ClinicalTrials.Gov NCT00461630 , Online)	The definition includes coronary death or revascularisation which are not only MI+Stroke - needs some adjustment
PROACTIVE	MI + Stroke + Any Death	1	0.6	0.7	0.5	0.8	0.4	(ClinicalTrials.Gov NCT02678676, Online)	Includes many more elements including amputation and procedures - needs a reduction for sure

Note that the interpretations here were given by one person "impersonating" several opinions. Yet after computation, a merged interpretation is created by weighting all those interpretations together in a way that best matches all the other data and assumptions added to the system with regards to the query used. The spread in expert interpretations also can be used do define possible bounds for the ground truth value - is it quite possible for an expert to have several opinions on what is possible in case variability is large. The assumption engine will find the best fit.

RESULTS

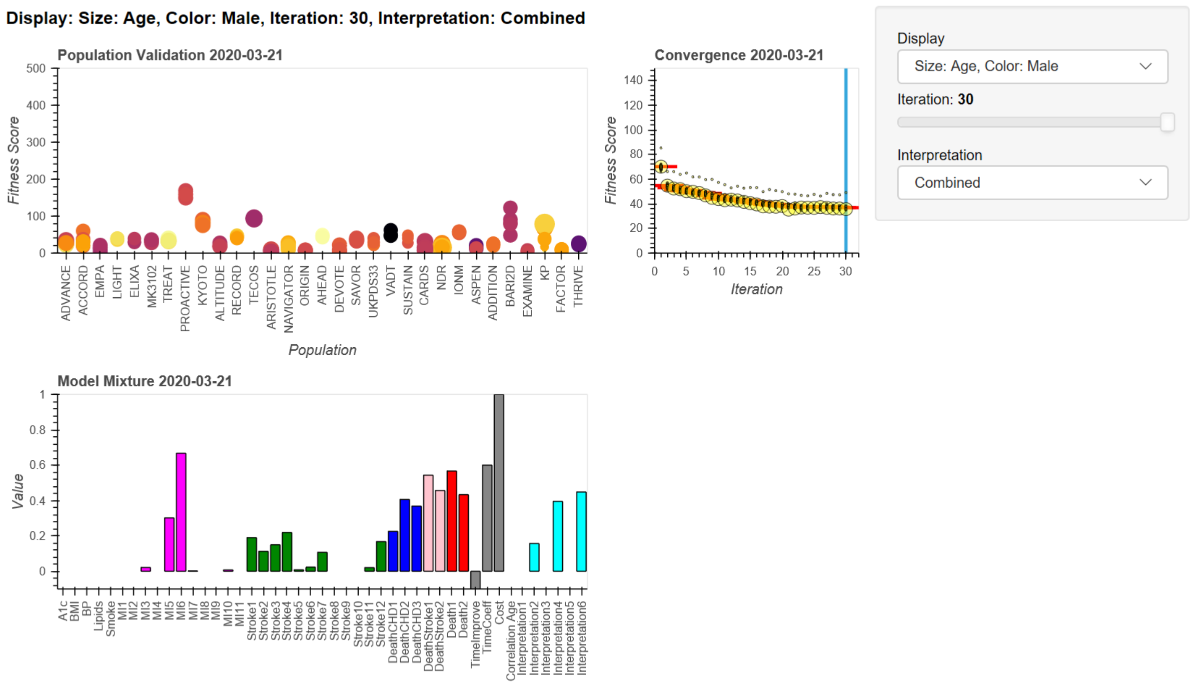
Simulation was conducted on a 64 core machine for 3 weeks. 30 optimization iterations were calculated to determine the most fitting model combination and the most fitting expert interpretation. When simulation started we already expected that one of the implemented risk equations that was shown to be misbehaving in the past would be eliminated by assumption engine. From past results it was known that the population we called PROACTIVE (ClinicalTrials.Gov NCT02678676, Online), since it was based on a previous trial enrollment with this acronym, was a severe outlier as can be see here (Barhak, 2019) . So we expected that Expert 1 interpretation will be rejected by the assumption engine. Recall that expert interpretation 1 simulates an expert that believes that the model outcomes are defined the same as the study outcomes - looking at the clinical trial definition of the outcome, we know this is not reasonable and in fact this may have been better if this trial was excluded from validation due to incompatibility. However, in this work it serves a purpose of showing how human interpretation can help explain things. The results generated do support our prior knowledge and MI equation 11 and expert interpretation 1 weights are both zero at the end of simulation as can be seen in figure 2.

The Reference Model Visualization was enhanced once more this year to use the most advanced HoloViz python technology to visualize the results interactively. Those interactive visualization allow hovering with the mouse over plot elements to get more information. To supplement this paper, some iterative visualization are available online at: (https://jacob-barhak.netlify.com/thereferencemodel/results_2020_03_21_visual_2020_03_23/CombinedPlot.html), the interactive visualization shows interactively what is shown in figure 2 statically as one snapshot and will take a long time to load as the file size is nearly100Mb - a good internet connection and strong machine are advised.

Figure 2 shows 3 plots: the top left plot represents clinical trials cohorts and their fitness. Each circle is a clinical trial and its color / size represent Age and proportion of Male and their height represents the fitness of model prediction to the observed outcomes of the clinical trial cohort. Fitness may include multiple outcomes associated with the study that are merged into one number, for the sake of simplicity think about it as simulation error measure for that cohort, defined by the query posed to the model. So a higher circle on the vertical axis, means that that cohort results cannot be explained well compared to a cohort that is represented by a lower circle. Ideally we want all circles to be as close to zero as possible, meaning that our ensemble model is very good. However, this is not realistic, since even observed clinical trial results have statistical variability. However, this plot is useful since it shows us what we can explain well computationally. In the future addressing issues that cause some cohorts to be predicted poorly, may improve fitness. So this result give a reference for comparison of our cumulative computational knowledge. The more information that can be absorbed Into the model the better we can see how well computers can explain and predict a phenomenon. The Reference Model is than important as a map for exploration of the ability of machines to comprehend medical knowledge.

The bottom plot in Figure 2 represents the weights that construct the best model. Each bar is associated with a certain equation, while equations that represent the same transitions have the same color. The last group of bars colored cyan is associated with the interpretations. It is clear that there is no bar for MI equation 11 and no bar for expert interpretation 1, meaning that those assumptions have been rejected by the assumption engine as not contributing to the most fitting model.

The Top right plot represents the convergence of the model in each simulation iteration. The overall fitness score, that is a weighted average of cohort fitness scores, is shown as big circles. The fitness of gradient components is shown as smaller circles. It can be seen how the simulation converges and stays more or less steady after 30 iterations. Since the simulation is Monte-Carlo based it is expected to see some fluctuations, yet the results show clear convergence. If we look at the last combined fitness score of ~36 out of 1000 and trying to best interpret the math, we can very loosely say that according to all the knowledge accumulated to date, and while making many simplification in result interpretations, we can predict outcomes on average with fitness of 3.6%. This is our current cumulative gap of computational knowledge and an improvement of 1.4% over the result on 2019 (Barhak, 2019).



Display: Size: Age, Color: Male, Iteration: 30, Interpretation: Combined

Display

Size: Age, Color: Male

Iteration: 30

Interpretation

Combined

Figure 2. The Reference Model Results

DISCUSSION

The Reference Model in about 8 years of development accumulated more computational knowledge than ever was reported to be accumulated by any diabetes CVD model. Not only it can absorb other models, assumptions, populations, it can now also include human interpretation. The ensemble model now allows automation of significant portions of the modeling process, processes that were once, and even today, done manually.

The Reference Model rise in capabilities by automation should also be contrasted against the decline in human modeling capabilities as reflected by the Mount Hood Diabetes challenge group. The Reference Model was initially created to imitate and improve some processes happening in validation challenges in 2010. In 2012, 2014 the human modeling groups participating, did not validate their results against previous year results while the ensemble model did validate against all previous populations - 8 in 2014. The Mount Hood challenge in 2014 only validated against one population and in 2016 no more populations were introduced for validation, while the ensemble model grew in its validation capabilities in these years while adding those to previous populations and reaching 9 population in 2016 and today stands on 30. The decline of the human modeling paradigm was very clear in 2016 Mount Hood Diabetes Challenge where human groups, including the author, were asked to recreate previous models without success by any team (Economics Modelling and Diabetes: The Mount Hood 2016 Challenge, Online). This alone proves that humans should not be performing repetitive modeling tasks that are better done by machines. However, human decline has reached a new low when some participants in the challenge decided to republish the 2016 challenge results while omitting results - humans can decide to do this, while machines do not remove data willingly. The Reference Model results were removed while it was the only model that has reproducibility tests build within it - see reproducibility section below. During the challenge and afterwards during the summary process the author has called multiple times for publication for code for reproducibility and the idea was not adopted by the human led group.

This decline in human modeling approach compared against rise in automation capabilities and accumulation of knowledge by machines happens in other aspects of our lives like driver-less car technologies that are slowly developing. However, despite machine automation rise, humans still have value and their opinions and needs should be collected by machines in proper manner. The machine automate tasks well, while humans should have a good interface to guide the machines to reach desired goals. The Reference Model now has proper interfaces for humans that fulfill the following roles: 1) Modelers can add new models/assumptions to our knowledge, 2) Data experts /

Bio statisticians can archive clinical trial data to be validated against 3) Medical experts can interpret clinical trial definitions. Using those interfaces and further improving automation and gathering of data, it would be possible to improve our model prediction accuracy in the future. At some point in time, machine prediction accuracy should become comparable to the average medical expert prediction - this phenomenon is already reported for other machine automated tasks (Laserson, 2018) . When this point is reached and validated, it may be possible to discuss government approval of deploying such technologies. In fact the government is already preparing towards such scenarios (FDA - SaMD, Online). Some prediction on when this machine takeover may happen can be found in (Barhak & Schertz, 2019). The good news are that deployment of machine based technologies is easy and fast compared to deployment of traditional medical knowledge that is accomplished by long cycles of training humans, recruitment, knowledge exchange, and retirement, that take years. Software deployment, even considering hurdles is much faster. So the time from policy approval to deployment is relatively fast, and human adoption will not be hard for technologies that proved themselves if human concerns are addressed.

Therefore the current effort should be in improving the ability of machines to predict and accumulate knowledge. The Reference Model is only one tool in this struggle and it shows that our cumulative computational capability still needs improvement. However, other technologies that help in accumulation of data and its standardization like (ClinicalUnitMapping.Com, Online) are already under development and will allow improving the knowledge accumulation pipeline.

REPRODUCIBILITY INFROMATION

Results presented in this paper were extracted from a simulation executed on a 64 core machine using Ubuntu 18.04.01 LTS with python 2.7.15 delivered by Anaconda with dask 0.19.1 supporting multi processing and MIST 0.94.6.0 as the simulation engine. The simulation results was archived under the file MIST_RefModel_2020_03_21_OPTIMIZE.zip. Formula Validation run to validate integrity was archived as MIST_RefModel_2020_01_02_FORMULA.zip. Visualization was generated on a notebook machine using Windows 10 x64 with Python 2.7.16, Bokeh: 1.4.0, Holoviews: 1.12.7, panel 0.8.0. The code/data files VisualExploration_2020_03_23.zip archives the visualization results.

ACKNOWLEDGEMENTS

Thanks to Deanna J. M. Isaman who first introduced the idea of accumulating clinical knowledge from public data. Many thanks to Philipp Rudiger, James Bednar, and Jean-Luc Stevens for developing HoloViz Visualization Tools.

REFERENCES

Barhak J., Isaman D.J.M., Ye W., Lee D. (2010), Chronic disease modeling and simulation software. Journal of Biomedical Informatics, Volume 43, Issue 5, October 2010, Pages 791-799, http://dx.doi.org/10.1016/j.jbi.2010.06.003

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

Barhak J. (2014). The Reference Model for Disease Progression – Data Quality Control. Monterey CA. Paper retrieved from: http://dl.acm.org/citation.cfm?id=2685666 Presentation retrieved from: http://sites.google.com/site/jacobbarhak/home/SummerSim2014_Upload_2014_07_06.pptx

Barhak J., Garrett A., (2014). Population Generation from Statistics Using Genetic Algorithms with MIST + INSPYRED. MODSIM World 2014, April 15 - 17, Hampton Roads Convention Center in Hampton, VA. Paper: http://sites.google.com/site/jacobbarhak/home/MODSIM2014_MIST_INSPYRED_Paper_Submit_2014_03_10.pdf Presentation: http://sites.google.com/site/jacobbarhak/home/MODSIM_World_2014_Submit_2014_04_11.pptx

Barhak J. (2015). The Reference Model uses Object Oriented Population Generation. SummerSim 2015. Chicago IL, USA. Paper retrieved from: http://dl.acm.org/citation.cfm?id=2874946 Presentation retrieved from: http://sites.google.com/site/jacobbarhak/home/SummerSim2015_Upload_2015_07_26.pptx

Barhak J., Garrett A., & Pruett W. A. (2016). Optimizing Model Combinations, MODSIM world, Virginia Beach, VA. Paper retrieved from: http://www.modsimworld.org/papers/2016/Optimizing_Model_Combinations.pdf Presentation: http://sites.google.com/site/jacobbarhak/home/MODSIM2016_Submit_2016_04_25.pptx

Barhak J. (2016), The Reference Model for Disease Progression Combines Disease Models. I/IITSEC 2016 28 Nov – 2 Dec Orlando Florida. Paper: http://www.iitsecdocs.com/volumes/2016 Presentation: http://sites.google.com/site/jacobbarhak/home/IITSEC2016_Upload_2016_11_05.pptx

Barhak J. (2017), The Reference Model Estimates Medical Practice Improvement in Diabetic Populations. SpringSim, April 23 –26, 2017, Virginia Beach Convention Center, Virginia Beach, Virginia, USA.

Barhak, J. (2019) The Reference Model is the most validated diabetes cardiovascular model known. MSM/IMAG meeting. IMAG Multiscale Modeling (MSM) Consortium Meeting March 6-7, 2019 @ NIH, Bethesda, MD . Poster: https://jacob-barhak.github.io/InteractivePoster_MSM_IMAG_2019.html

Barhak J. (2020), The Reference Model Accumulates Knowledge With Human Interpretation. Interagency Modeling and Analysis Group - IMAG wiki - MODELS, TOOLS & DATABASES Uploaded 16 March 2020. Poster: https://jacob-barhak.github.io/Poster_MSM_IMAG_2020.html

Jacob Barhak, Joshua Schertz (2019). Standardizing Clinical Data with Python . PyCon Israel 3-5 June 2019, Video: https://youtu.be/vDXyCb60L5s Presentation: https://jacob-barhak.github.io/Presentation_PyConIsrael2019.html

Bokeh, (Online). https://docs.bokeh.org/en/latest/index.html

Holoviz, (Online). https://holoviz.org/index.html

Clarke P.M., Gray A.M., Briggs A., Farmer A.J., Fenn P., Stevens R.J., Matthews D. R . Stratton. I. M., Holman R. R., &UK Prospective Diabetes Study (UKDPS) Group (2004). A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68). Diabetologia, 47(10),1747-59. http://dx.doi.org/10.1007/s00125-004-1527-z

ClinicalTrials.gov - NCT00379769: Rosiglitazone Evaluated for Cardiac Outcomes and Regulation of Glycaemia in Diabetes (RECORD) (Online) https://clinicaltrials.gov/ct2/show/results/NCT00379769?view=results

ClinicalTrials.gov - NCT00461630: Treatment of HDL to Reduce the Incidence of Vascular Events HPS2-THRIVE (HPS2-THRIVE) (Online) https://clinicaltrials.gov/ct2/show/results/NCT00461630?view=results

ClinicalTrials.gov - NCT02678676: Rosiglitazone Evaluated for Cardiac Outcomes and Regulation of Glycaemia in Diabetes (RECORD) (Online) https://clinicaltrials.gov/ct2/show/results/NCT00379769?view=results

ClinicalUnitMapping.Com (Online): https://clinicalunitmapping.com/

Eddy D.M. , Schlessinger L. (2003), Validation of the Archimedes Diabetes Model, Diabetes Care 2003 Nov; 26(11): 3102-3110. https://doi.org/10.2337/diacare.26.11.3102

Gary T., Mingle D., Yenamandra A.(2016) The Evolving Definition of Sepsis. arXiv:1609.07214v1. https://arxiv.org/ftp/arxiv/papers/1609/1609.07214.pdf

Griffin S. J. Borch-Johnsen K., Davies M.J., Khunti K., Rutten G., Sandbæk A., (2011). Effect of early intensive multifactorial therapy on 5-year cardiovascular outcomes in individuals with type 2 diabetes detected by screening (ADDITION-Europe): a cluster-randomised trial. The Lancet, VOLUME 378, ISSUE 9786, P156-167, https://doi.org/10.1016/S0140-6736(11)60698-3

Laserson J., Lantsman C. D., Cohen-Sfady M., Tamir I.,Goz E. Brestel C., Bar S., Atar M, Elnekave E. (2018). TextRay: Mining Clinical Reports to Gain a Broad Understanding of Chest X-rays. arXiv:1806.02121v1 , https://arxiv.org/abs/1806.02121

Leff, H. S., Hughes, D., Chow, C., Noyes, S., & Ostrow, L. (2009). A Mental Health Allocation and Planning Simulation Model: A Mental Health Planner's Perspective. In Y. Yuehwern (Ed.), Handbook of Healthcare Delivery Systems. http://www.hsri.org/files/Mental%20Health%20Allocation%20and%20Planning%20Simulation%20Model-Final-PDFversion.pdf

Hayes A.J., Leal J., Gray A.M., Holman R.R., & Clarke P.M. (2013). UKPDS outcomes model 2: a new version of a model to simulate lifetime health outcomes of patients with type 2 diabetes mellitus using data from the 30 year United Kingdom Prospective Diabetes Study: UKPDS 82. Diabetologia, 56(9), 1925-33. http://dx.doi.org/10.1007/s00125-013-2940-y

Palmer A.J., & The Mount Hood 5 Modeling Group (2013). Computer Modeling of Diabetes and Its Complications: A Report on the Fifth Mount Hood Challenge Meeting, Value in Health, 16(4), 670-685. http://dx.doi.org/10.1016/j.jval.2013.01.002

FDA - SaMD (Online) - Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback (Online). https://www.regulations.gov/document?D=FDA-2019-N-1185-0001

Stevens R., Kothari V., Adler A., Stratton I. (2001), The UKPDS risk engine: A model for the risk of coronary heart disease in type II diabetes UKPDS 56. Clin Science, 2001; 101: 671-679.

The Mount Hood 4 Modeling Group (2007). Computer Modeling of Diabetes and Its Complications, A report on the Fourth Mount Hood Challenge Meeting. Diabetes Care, (30), 1638–1646. http://dx.doi.org/10.2337/dc07-9919

Economics Modelling and Diabetes: The Mount Hood 2016 Challenge (Online). https://docs.wixstatic.com/ugd/4e5824_0964b3878cab490da965052ac6965145.pdf

UK Prospective Diabetes Study UKPDS Group (1998). Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes UKPDS 33. Lancet, 1998; 352: pp.837-853.

Wilson P. W. F., D'Agostino R.B., Levy D., Belanger A.M., Silbershatz H., Kannel W. B. (1998), Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation 1998;97;1837-1847, https://doi.org/10.1161/01.CIR.97.18.1837

Wentowski C., Mewada N., Nielsen N. D. (2019) Sepsis in 2018: a review . Anaesthesia & Intensive Care Medicine Volume 20, Issue 1, Pages 6–13. https://doi.org/10.1016/j.mpaic.2018.11.009

Wikipedia, Delphi method, (Online) https://en.wikipedia.org/wiki/Delphi_method