A Proposal for Standardized Evaluation of Epidemiological Models

(A position paper)

Roni Rosenfeld, PhD[+] *, John Grefenstette, PhD[++], Don Burke, MD[++]
[+]Carnegie Mellon University, [++]University of Pittsburgh

Extended Abstract

Public Health (PH) officials make many decisions during both emergency and non-emergency periods: procurement, vaccine formulation, resource pre-positioning, surveillance intensity, school closure recommendations, etc. These decisions are made in an uncertain environment, and would benefit greatly from accurate estimates of the likelihood of different outcomes. For example, to decide:

- how much to invest in surveillance of a particular animal virus, it's important to know whether its probability of jumping to humans over the next X years is 10%, 1% or 0.1%.
- whether to purchase new respirators, it's important to know how likely the impending epidemic is to exceed the current capacity.
- how aggressively to use a given drug, it's important to know how likely resistance is to emerge within, say, a year, under the different possible drug usage policies.
- whether to recommend school closure, it's important to know the likely epidemic peak, timing and attack rate under the different options.

Public Health officials need quantified, evidence-based likelihood assessments: not whether or not something will happen, but how likely different outcomes are. Not only what is about to happen, but also what is likely to happen under different intervention scenarios.

Different models have been built to address these questions. Recently there has been much progress in model sophistication and realism. But different models are based on different assumptions, and can produce competing predictions or forecasts. PH officials currently have no principled way of telling (1) how accurate a given model is; (2) which model is most accurate; or (3) how to best combine predictions or forecasts from the different models. In the words of one veteran modeler, "testing models by prediction has not yet become standard practice in epidemiological modeling" [1, p. 393].

In this paper, we argue for setting *forecasting accuracy* as an explicit goal of epidemiological modeling. Specifically, *we propose a public, ongoing evaluation program for measuring the forecasting accuracy of epidemiological models*. The program should be designed by all stakeholders: PH decision makers, practitioners, funders and researchers. The evaluation should be based primarily on forecasting of events to occur in the coming 6-12 months, such as the dominant strain in the next epidemic wave, the number of documented animal-to-human transmissions, number of respirators used, etc. Based on the first author's experience with designing standardized evaluations in other fields, we recommend the following process:

1) Convene a planning workshop to decide on focus and general guidelines for the evaluation.
2) Establish a governing body to oversee the evaluations on an ongoing basis.
3) Meet annually to analyze the results and revise the evaluation program.

* Corresponding author, roni@cmu.edu

Standardized Evaluations as Catalyst for Rapid Progress

In the late 1980s, DARPA and related DoD agencies initiated community-wide standardized evaluations of automatic speech recognition (ASR) technology for their grantees and other interested R&D groups [2,3]. Dramatic improvements in ASR capabilities soon followed (figure 1). Other factors have also contributed to the dramatic improvements in the state-of-the-art of speech technologies of the 1980s and 1990s. But without standardized evaluations it would not have even been possible to know what the state of the art was, nor to track its progress over time.

The idea caught on fast. Within a few years, standardized evaluations were established in machine translation, information retrieval, question answering (leading to Watson, [4]) and a variety of other language technologies (for a full list, see [5]). The not-for-profit X Prize Foundation ([6]) has been successfully inducing technological competitions with cash prizes since 1994. In computational biology, a standardized evaluation program for protein folding (CASP, [7]) was introduced in 1994, and was followed in 2000 by standardized prediction of protein-protein docking (CAPRI, [8]). The DARPA Grand Challenge for driverless vehicles (2004—2007, [9]) was extremely successful. Standardized evaluations in Machine Learning are now common, and have recently become famous with the Netflix Prize (2006—2009, [10]).

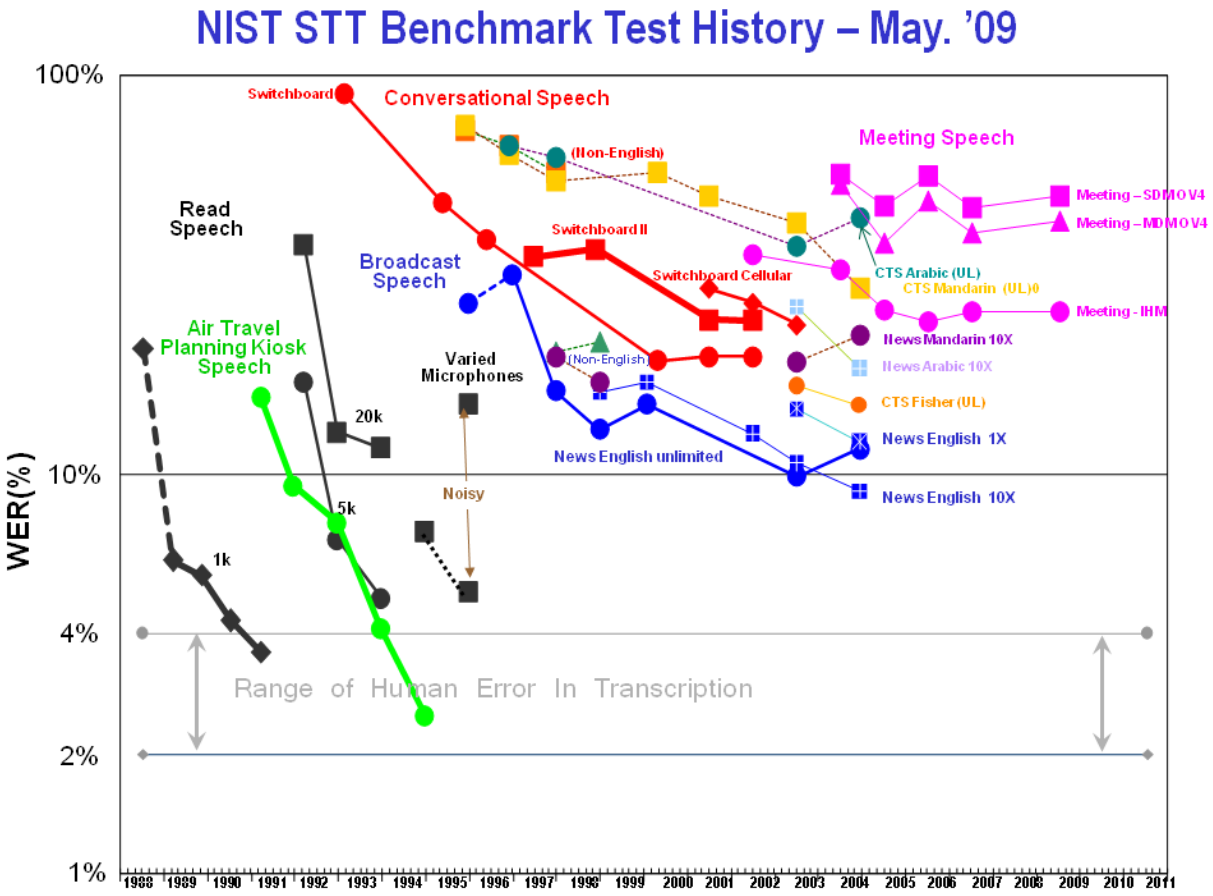

Figure 1: Dramatic progress in speech recognition accuracy under the standardized evaluation paradigm. Each Word Error Rate (WER) curve corresponds to a particular, well defined standardized evaluation task. From [2].

We can use the experience and lessons from these standardized evaluations to design an evaluation program for epidemiological models. Such a program will (1) measure for the first time the state of the art, (2) allow us to track progress over time, and (3) induce the epidemiological research community to devote effort to improving forecasting accuracy, which is where the greatest practical need lies.

Evaluating the Forecasting Accuracy of Epidemiological Models

We have been using the terms "prediction" and "forecast", and would now like to define them, following [13]. By "prediction" we mean an assertion that a certain outcome event will occur. Once it becomes clear whether or not the

event in question had in fact occurred, the prediction can be said to have been proven *right* or *wrong*. In contrast, by "forecasting" we mean the assignment of a probability distribution[1] over the range of all possible, mutually exclusive outcome events. Once the actual outcome is known, the probability assigned to it by a given forecast attests to the *accuracy* of that forecast. Thus, a forecast cannot be said to be "right" or "wrong", but only to be more or less accurate. For more details, and for why we prefer forecasting to prediction, see Appendix I.

The ultimate test of forecasting accuracy is to actually forecast the future, rather than forecasting held-out past data ("post-casting" or back-testing). This is especially true in public evaluations involving many different approaches. Luckily, a rich stream of epidemiological surveillance data is continuously collected nationally and worldwide. Even rare events are recorded quite frequently (since the number of different possible rare events is very large). Consequently, a large number of different types of testable forecasts can be made about the near future. A partial sample includes forecasts of:

- The number of reported cases of a disease in a given period of time (rather than the true incidence of the disease, which is hard or impossible to measure)
- The peak demand for respirators in a given metropolitan area during a given Influenza epidemic season.
- The detection of new animal-origin viral strains in humans.
- The detection of human-to-human transmission of an animal-origin strain, optionally conditioned on such a strain having been previously detected in humans (see discussion of conditional forecasting below).
- Disease-specific mortality figures, conditioned on each of several different interventions.
- The date of onset of the next seasonal epidemic (suitably defined in terms of measurable quantities).
- The dominant flu type, subtype and strain in an upcoming epidemic, and the degree of dominance.
- The clinical severity of an upcoming epidemic (suitably defined in terms of measurable quantities).
- The future geographic distribution of a disease (e.g. new outbreaks of polio)
- The impact of an upcoming vaccination campaign (suitably conditioned on the campaign's reportable outcomes, e.g. number of vaccines administered)
- The timing, location and/or extent of the reported emergence of drug resistance.
- Age-stratified levels of seropositivity to a given antigen, ahead of a planned serosurvey.

Any of these forecasts can be further refined by specifying the relevant time period, the date by which the forecast must be committed to, and/or the data on which the forecast may be based.

Any subset of the above forecasts can be combined to measure the overall forecasting accuracy of models. When a model is used to forecast the outcomes of many different events, its overall forecasting accuracy can be estimated reliably, even if most or all of the events considered are rare. This is done using the concept of cross entropy, as described in appendix I. Different models can then be compared objectively and quantitatively.

Setting Up the Evaluation Program

The suggestions below are based on the first author's experience with the design and implementation of standardized evaluations in other fields. No doubt they will need to be adjusted to the specific needs and realities of epidemiological research and public health practice.

We should convene a **planning workshop** to jump-start the process. It should be attended by senior people from the following constituencies:

- PH planners, practitioners and decision makers: CDC, BARDA and other HHS agencies, DOD agencies, state and local PH agencies.
- Funders (both public and private) of epidemiological research.
- epidemiological modelers, and members of the V&V community.

The most important decisions to be made during this phase involve **which epidemiological events should be included** in the standardized evaluation, because that choice will tend to drive much of subsequent research. Given the vast number of possible forecasts of relevance to public health, a sample should be chosen to best reflect the needs of public health decision makers. Involvement by the latter is crucial during this step, to maximize the utility of the forecasting technology that will be developed. The best forecasts to focus on would be those that (1) are based on obtainable data and

[1] or a probability density function in the case of a continuous-valued outcome.

(2) will contribute to informed decisions that could lead to significantly improved public health outcomes. The choice should also reflect the interests and capabilities of the modeling community.

For a forecast to be included in the evaluation, its eventual outcome must be incontrovertible. For examples of how to define these properly, see commercial prediction markets, e.g. Intrade [11].

The workshop should produce written guidelines for the evaluation program. It should also nominate a **governing body**: a small standing committee representing the stakeholders above to design, oversee and iteratively refine the evaluation program. The committee will need to meet regularly by conference call, and occasionally face-to-face.

The evaluation plan should include **multiple evaluation tracks.** Each participating team could choose to participate in any subset of these tracks but must announce its decision ahead of time.

All evaluation participants should submit their forecasts to a **third party evaluator**, which will timestamp them and carry out the evaluation and comparative statistical analysis. The evaluation methodology will be published in full detail in advance. Note that no programs or algorithms need to be submitted by participants (unless otherwise agreed) – only numerical forecasts.

Annual **"post-mortem" meetings** should be held to discuss the results, exchange ideas and set direction for future evaluations.

Data sharing policy: Models and methods can best be compared if they are based on the same data. On the other hand, our goal is to maximize forecasting accuracy, and many would-be participants may not be willing or able to share all their data. Data sharing policy needs to be negotiated in advance. One solution is to base some evaluation tracks on shared data, and allow others to be based on additional, private data. Government-funded programs may impose their own requirements regarding data sharing.

Method sharing policy: Science and technology best progress under complete sharing of ideas. At the same time, commercial and other private actors should be encouraged to participate in these evaluations as well. The required degree of post-evaluation sharing should be negotiated in advance. Academic participants may wish to release their models and methods down to the code level. A somewhat less rigorous standard of sharing would be what is currently required in scientific publications, namely enough detail to enable anyone "versed in the art" to reproduce the results with some effort. Government-funded programs may impose their own requirements regarding method sharing.

Bragging Rights: groups that perform well may understandably want to advertize that fact. But the comparative analysis can sometimes be subtle. To avoid a media war, binding policy should be carefully worked out in advance. One solution is for participants to agree to avoid any public announcements regarding their performance, relying instead on the Evaluator's website and reports.

Funding: There must be separate funds explicitly allocated to establishing and running the standardized evaluations. These will cover the third party Evaluator, the annual "post mortem" meetings, the work of the standing committee, and perhaps also shared data collection. Beyond this, funders may wish to designate research funds specifically for development of forecasting technologies, so it is clear to the funded researchers that this is what they are asked to do, and to funders what their money is invested in.

Possible Objections

Below are a few possible objections to standardized evaluations, together with rebuttals or amelioration strategies.

Research focus might shift away from science towards technology. This may well happen, but is not necessarily a bad thing. Budget allocation should reflect the relative societal need for science vs. technology in this area, and the funders' priorities. Arguably, in epidemiological research, the pendulum is currently swung too far away from technology. Enlightened funders may choose to fund a diverse portfolio of science (which they are now doing) and technology (as measured by forecasting accuracy). A standardized evaluation program will allow funders and others to see their grantees' current forecasting capabilities and make an informed decision.

Research groups will "study to the test". The evaluation must be carefully designed to test the forecasting capabilities that are most needed, and then "studying to the test" would be a good thing. Ultimately, successful forecasting methods, necessarily based on sound scientific and mathematical principles, will better generalize to tests they were not explicitly designed for.

Small players could not afford to compete with large research teams. They could specialize and compete in one of the tracks. They could also collaborate with a larger group to provide a defined sub-contribution. Alternatively, they could compete for scientifically- (rather than technologically-) oriented funds. Without standardized evaluations, the potential contributions of small players to forecasting technology cannot be assessed at all.

The evaluations will create a competitive atmosphere among different research groups, reducing the level of sharing and disclosure. In the Automatic Speech Recognition community, exactly the opposite has happened. The shared tasks and metrics resulted in an easier exchange of information. The extent of method- and data-sharing needs to be worked out in advance and made a condition of participation. As participation becomes encouraged by funding agencies, this will tend to increase the level of sharing. The shared evaluation will also provide an impetus for the shared identification, collection and processing of data that is collectively deemed as most useful for making epidemiological forecasts. As an example, the speech recognition evaluations were an impetus for the establishment of the extremely successful Linguistic Data Consortium ([12]).

Appendix I: A Mathematical Primer on Forecasting

Predicting quantity vs. predicting events: When predicting a numerical quantity, like the attack rate of an epidemic or peak respirator usage, predictions offered by different models can be compared, after the true outcome is known, simply by comparing their respective errors.

But when modeling categorical events (one of several mutually exclusive outcomes), and especially when some of the outcomes are rare, a more general framework, *forecasting*, is needed to compare the accuracy of different models.

As an example, consider predicting the dominant Influenza H3N2 antigenic strain in the coming season (which is important for deciding on vaccine formulation for that season). Assume that three different models (M1, M2, M3) participate in the competition, each providing its respective estimate of the probabilities of each of the three outcomes to be considered (Table 1).

Next season's dominant H3N2 strain:	**Probability acc. to Model M1**	**Probability acc. to Model M2**	**Probability acc. to Model M3**
A/California/07/2004	90%	55%	70%
A/Korea/770/2002	5%	40%	29%
None of the above	5%	5%	1%
Total	100%	100%	100%

Table 1: Example of comparing the predictions and forecasts of 3 models

If asked to predict which strain will dominate, all three models will choose "A/California/07/2004/", since this is the most likely outcome according to all of them. Note that, w*hether or not this turns out to be correct, nothing will be learned about the relative accuracy of the models.*

Instead, we ask the models to provide their complete probability distribution estimate (the column of three numbers above, which always must sum to 100%). After the actual outcome becomes known, we compare the models on the log-likelihood they assign to the event that actually happened. Thus, if the dominant strain turned out to be "A/California/07/2004/", M1 will score the best (log(0.9)), followed by M3 (log(0.7)) and then M2 (log(0.55)), although the differences between the scores will be relatively small. If, on the other hand, "none of the above" turned out to be the correct answer, models M1 and M2 will be scored the same (log(0.05)), but M3 will be assigned a significantly worse score (log(0.01), a difference of log(5)). This is appropriate, because M3 significantly (and potentially dangerously) under-estimated that possibility.

When a model is used to forecast the outcomes of many different events, its overall forecasting accuracy is estimated by summing the log-likelihoods it assigns to all the outcomes that actually happened. This is a commonly used and well understood measure. In Statistics, it is known as "log-likelihood of new observations". In Information theory, it is known as the cross-entropy of the true distribution of events relative to the model's estimate. In language technology, a variant of this measure goes by the name "perplexity".

The log-likelihood method is very general. Even a prediction of quantity can be converted to this framework by providing, instead of a single number, a probability distribution (or a probability density function) over all possible numerical outcomes. Therefore, forecasting of quantities and of discrete events can be combined when estimating forecasting accuracy and when comparing different models.

Forecasting rare events: The log-likelihood framework can be used to estimate the accuracy of forecasting common events, as well as moderately rare ones (e.g. those with probability ≈ 10%). For example, to estimate the accuracy of forecasting the occurrence of a new Influenza subtype, one does not need to wait until enough such events have happened. Instead, the models can be consulted every year, in fact every month, and two such models can begin to be compared even when no new reassortants have occurred. By combining the assessments of many such events, each tending to occur only 10% of the time, we get a more reliable estimate of overall accuracy.

Assessing the accuracy of forecasting much rarer events (e.g. those with probability ≈ 1%) is much harder. One strategy would be to break the event of interest into a set of successive intermediate steps, each having a somewhat higher probability, and to have the models forecast each step conditional on the one before it. For example, the event "a new zoonotic pathogen will invade the human population" can be broken into the following verifiable sub-events: measured spread in the original host population, documented animal-to-human transmission, documented human-to-human transmission, and observed human population spread.

Conditional forecasts: Conditional forecasts are useful too, but the conditioning events must become known and incontrovertible before the forecast event. Thus, we can forecast disease incidence on a particular week given the observed incidence k weeks before, for any value of k. Forecasts can also be made conditional of each one of several possible interventions. Only the forecast based on the intervention that actually took place (or the conditioning event that actually transpired) will be assessed.

Confidence assessment: It is also important to know how confident a given model is in its forecast – its confidence level. This is the "know when you don't know" problem, and corresponds to a gambler's decision to sometimes "sit out" a particular bet, rather than bet anything on it. It is especially important to know when the uncertainty is due to our lack of knowledge or understanding, vs. when it is due to the intrinsic uncertainty of the situation. The accuracy of a model's self-confidence can be estimated in much the same way, and can be made part of the standardized evaluation, as has been done in the Automatic Speech Recognition competitions starting in the mid 1990's [2].

---

[1] Predictability and preparedness in influenza control. Smith DJ. *Science*. 2006 Apr 21;312(5772):392-4. PMID: 16627736.
[2] http://itl.nist.gov/iad/mig/publications/ASRhistory/index.html
[3] http://itl.nist.gov/iad/mig/tests/rt/ASRhistory/pdf/NIST_benchmark_ASRtests_2003.pdf
[4] http://www-03.ibm.com/innovation/us/watson/.
[5] http://www.itl.nist.gov/iad/mig/tests/
[6] http://www.xprize.org/.
[7] http://predictioncenter.org/
[8] http://www.ebi.ac.uk/msd-srv/capri/.
[9] http://www.darpagrandchallenge.com/.
[10] http://www.netflixprize.com/.
[11] http://www.intrade.com/.
[12] http://www.ldc.upenn.edu/.
[13] The Signal and the Noise: Why So Many Predictions Fail -- but Some Don't. Nate Silver. Penguin Press, 2012.