Five Minute Science Lesson: Multiple Testing Strategies in Two Different Contexts (SARS-CoV-2 and Polygraph)

Raymond Nelson

A basic purpose of any scientific test is to quantify, classify or predict a phenomena of interest, sometimes referred to as an *unknown parameter*, that cannot be subject to perfect deterministic observation or direct physical measurement. The basic procedure for any test of a test is to obtain some data, often referred to as a *sample*, that can be used to calculate a *statistical classifier* – using some form of statistical likelihood function (reference data or reference distribution) along with a structured process or rule to classify the result of a scientific test or experiment as either positive or negative. For example: the basic decision rule in the frequentist statistical tradition is this: $p < a = sig$. Test samples can be in the form of a *physical sample*, such as obtained by via nasopharyngeal swab or blood drawn by a phlebotomist medical technician in the case of medical tests, and can also be in the form of recorded *stimulus-and-response* trials for social/behavioral tests such as a polygraph test.

Regardless of the type of test, sample data is not itself the unknown parameter or phenomena of interest, but is a *proxy* that is correlated with the phenomena of interest to a sufficient degree that it can be useful in making statistical inferences about the unknown parameter of interest. Whereas medical/epidemiological tests, which make of physical samples, can make use of a single data source, social/behavior tests – including psychological tests and actuarial risk measures – will often make use of multiple sources of information from which response features can be extracted and combined.

All scientific tests are fundamentally probabilistic and for this reason are not expected to be infallible – they are expected to quantify the probabilistic strength or margin of uncertainty associated with a test result or conclusion. When a test is used to quantify, vs. classify, an unknown parameter, the statistical information will attempt to describe the statistical likeli-

hood that the numerical value of the unknown parameter exists within a certain range. Many tests are intended only to classify (prediction can be thought of as a form of classification). Ideally, though not always, a statistical classifier will also provide information about the practical strength of the information or conclusion, or the margin of uncertainty.

The statistical classifier for some scientific tests is abstracted from the practical context to a degree that, although it can be used to classify a test result, there may not be convenient or intuitive relationship between the test statistic and practical considerations such as the actual likelihood of correct or incorrect test outcomes – often referred to TP or *sensitivity* and TN or *specificity* and FP and FN rates. P-values – used to estimate random measurement error – are an example of this; they can be used to classify the results of a scientific test or experiment as *statistically significant* or *not significant* according to an alpha tolerance level, but do not provide information about practical likelihoods associated with the classification. Practical outcomes are most often described empirically as to the sensitivity, specificity and FP or FN error rates that are observed at selected numerical or statistical decision thresholds. Even more practical outcomes can be achieved using Bayesian or *a-posteriori* methods that take into consideration both a test statistic and prior information.

Regardless of whether a test is medical/epidemiological or social/behavioral/psychological or actuarial, the basic concepts of scientific testing are similar. Also similar are the types of questions and strategies that developers of scientific tests will take into consideration when validating a test method. Another similarity is that testing is expensive, in terms of financial costs, human activity and time. When it is necessary to conduct a large volume of tests, efficiency – including time, physical resources and human activity – the need to maximize available resources can become an important consideration. For example, *how to test a large cohort of public safety applicants for their history of involvement in multiple possible behavioral issues that make signal their unsuitability for positions of public trust? Or, how to test the population of a large city for SARS-CoV-2 in attempt to isolate and contain the spread of disease?*

Multiple testing is a common strategy that can be used to increase testing efficiency. Use of multiple testing strategies can be observed in different testing contexts – including multiple issue polygraph testing, and also in the form of pooled testing for SARS-CoV-2, the novel coronavirus responsible for the COVID-19 pandemic. Multiple testing, in this usage, refers to the evaluation of multiple targets in a single analysis. In the SARS-CoV-2 context multiple testing strategies are referred to as *pooled* testing, wherein multiple samples are pooled together for analysis. Although other countries have already made use of pooled testing strategies, in the U.S. the FDA and CDC have only recently issued

guidance on the development and validation of these methods for diagnostic and screening tests necessitated by SARS-CoV-2 and COVID-19.

According to the CDC website:

> Diagnostic testing for SARS-CoV-2 is intended to identify occurrence at the individual level and is performed when there is a reason to suspect that an individual may be infected, such as having symptoms or suspected recent exposure, or to determine resolution of infection. Examples of diagnostic testing include testing symptomatic individuals who present to their healthcare provider, testing individuals through contact tracing efforts, testing individuals who indicate that they were exposed to someone with a confirmed or suspected case of coronavirus disease 2019 (COVID-19), and testing individuals present at an event where an attendee was later confirmed to have COVID-19.

The CDC website also provides information to differentiate diagnostic tests from screening tests:

> Screening tests for SARS-CoV-2 are intended to identify occurrence at the individual level even if there is no reason to suspect infection—e.g., there is no known exposure. This includes, but is not limited to, screening of non-symptomatic individuals without known exposure with the intent of making decisions based on the test results. Screening tests are intended to identify infected individuals without, or prior to development of, symptoms who may be contagious so that measures can be taken to prevent further transmission. Examples of screening include testing plans developed by a workplace to test its employees, and testing plans developed by a school to test its students, faculty, and staff. In both examples, the intent is to use the screening testing results to determine who may return and the protective measures that will be taken.

The general concept of diagnostic and screening tests is essentially identical to that describe in the APA Standards of Practice.

> 1.1.5 Diagnostic examination: An event-specific evidentiary or investigative polygraph examination conducted to assist in determining the veracity of an examinee regarding his or her knowledge of or involvement in a reported issue or allegation. Diagnostic examinations may address a single aspect or multiple-facts of an event.
>
> 1.1.6 Screening examination: A polygraph examination conducted in the absence of a reported incident or allegation. Screening examinations may be conducted as single issue or multiple issue exams.

Of importance here is that diagnostic tests are conducted in response to a known problem – an incident or allegation in the polygraph context, and disease symptoms or exposure in medicine and epidemiology. A tempting and easy mistake, for many, will be to conflate the two dichotomies: *diagnostic vs screening* and *single vs multiple* testing strategies. Administrative professionals and field practitioners who correctly understand these differences will be more apt to develop and implement testing strategies and policies that achieve their objectives.

In the polygraph context multiple testing strategies are commonly referred to as multiple-issue tests, and are sometimes referred to as multiple-facet tests – with only difference being whether a polygraph is intended for diagnostic or screening purposes. For multiple issue polygraphs the test stimuli are evaluated with an assumption of independent criterion variance. As an example: polygraph target issues for applicant screening of public safety employees can include, one's behavioral history with illegal drugs, commission of unreported serious crimes, domestic or intimate partner abuse, sexual assault, and hate crimes or social intolerance. It is conceivable that a person may have engaged in none, some, or all of these different types of behaviors.

A multiple testing strategy provides the advantage of increasing the sensitivity of the polygraph screening test to a wider range of behavioral concerns and is considered a more efficient use of time and other resources rather than attempting to investigate these different behaviors in separate examinations. A disadvantage of the multiple issue polygraph is a potential for reduced specificity, and precision. The heuristic for classification of multiple issue polygraph results is any-or-all, where a test result is classified as positive if any target question has produced a positive result, and is classified as negative if all test questions have produced negative results. (Also note that there is no known empirical advantage of a series of single issue exams compared to a multiple issue exam. To the degree that testing errors are a function of random measurement error, a series of single issue tests may be subject to multiplicity effects somewhat similar to those of a multiple-issue exam.)

Positive results from a multiple-issue polygraph may or may not indicate the exact area of problem behavior, and for this reason may result in additional testing of an applicant – depending on the size of the applicant group, level of interest in the individual, resources, risks, and other factors. It is also possible that an applicant may simply be adjusted, reduced or eliminated within the priority or hierarchy of available applicants. Evidence-based polygraph field practice standards do not permit examiners to render both positive and negative classifications from the same examination – including when the examination questions are developed with an assumption of independent criterion variance – because doing so would damage test accuracy (potentially crea-

ting a context for both FP and FN result in the same exam).

The FDA website provides additional guidance of for developers of pooled or multiple testing methods for SARS-CoV-2, with description of two different methods of combining multiple test samples (aliquot or partial media pooling and media-swab pooling):

> Generally, FDA recommends validating your test with either pooling approach in a way that preserves the sensitivity of your test as much as possible; that is, it is preferable to use an approach where all specimens identified as positive when tested individually are also identified as positive when tested using the pooled testing approach. However, a decrease in performance is likely with pooling strategies, due to dilution of the primary clinical sample. As discussed in the templates, since, sample pooling will greatly increase the number of individuals that can be tested using existing resources, a small reduction in sensitivity may be acceptable depending on the pooling efficiency and other mitigations in place. Therefore, FDA generally recommends that, after pooling, test performance includes ≥85% percent positive agreement (PPA) when compared with the same test performed on individual samples. Additional limitations, such as considering negative results from pooled samples to be presumptive negatives, may be recommended based on the patient population included in your clinical evaluation and the performance data submitted in your EUA [emergency use authorization] request.

The preceding paragraph is instructive for several reasons. Firstly, it acknowledges that multiple testing strategies can sometimes lead to a reduction of test sensitivity, and that care must be taken to avoid this. In the SARS-CoV-2 context test sensitivity – the ability of test to detect or identify the unknown phenomena of interest when it is present – is the metric of primary interest. In other contexts, it is possible that other metrics may be prioritized; such as test specificity – the ability of a test to correctly determine the absence of problem of interest. Pooled testing of SARS-CoV-2 samples differs somewhat from the polygraph example (involving multiple behavioral target issues and a single individual) in that it involves taking a portion of sample data for multiple individuals and conducting the analysis on a single combined sample. Classification of pooled test results of SARS-CoV-2 is similar to the classification of multiple issue polygraph results in that the entire pooled sample will be classified as either positive or negative – it is not possible to achieve both positive and negative results within a single analysis.

Importantly, this (FDA) paragraph illustrates that the selection of a testing strategy is inherently probabilistic and is always done with consideration for an ability to understand and interpret both

the scientific/probabilistic and practical meaning of the information from the test result. In this (SARS-CoV-2) example, there is a clearly stated objective that the concordance – referred to as *percent positive agreement* (PPA) – is intended to achieve an 85% threshold when comparing the results of multiple and single testing approaches.

A potential hazard of pooled testing of SARS-CoV-2 samples is that combining of samples may induce a reduction of test sensitivity as a function of the dilution of the signal of interest in the combined samples. Imagine, for example, a combined sample of 20 persons of whom only 1 is positive. It may be possible to dilute the viral load to a degree that falls below the cutscore or threshold for positive classification. This need to understand the test sensitivity rate is the reason pooled testing strategies must be subject to validation in addition to the validation of the analysis method itself. Approval of a pooled or multiple testing strategy should only be considered when acceptable test accuracy metrics can be maintained.

An example of this in the polygraph context is that numerical cutscores for positive classification of ESS-M subtotals are calculated without statistical correction for multiple issue tests. Use of a statistical correction for these classifications would reduce the incidence of FP errors but would do so at a cost of reduced test sensitivity and increased FN errors. For single issue polygraphs, for which overall precision is often an intended objective, it is expected that most classifications will be made using the total score, and for this reason the calculation of ESS-M numerical cutscores for positive classification of subtotal scores includes a statistical correction. For single issue exams there is no loss of sensitivity due to reliance on the total score. For these exams the use of subtotals with statistical correction can actually increase test sensitivity without a corresponding increase FP errors.

Another interesting aspect of the CDC website information is that information is also available on surveillance testing:

> Surveillance for SARS-CoV-2 includes ongoing systematic activities, including collection, analysis, and interpretation of health-related data that are essential to planning, implementing, and evaluating public health practice. Surveillance testing is generally used to monitor for a community- or population-level occurrence, such as an infectious disease outbreak, or to characterize the occurrence once detected, such as looking at the incidence and prevalence of the occurrence. Surveillance testing is used to gain information at a population level, rather than an individual level, and results of surveillance testing can be returned in aggregate to the requesting institution. Surveillance testing may sample a certain percentage of a specific population to monitor for

increasing or decreasing prevalence and to determine the population effect from community interventions, such as social distancing. An example of surveillance testing is a plan developed by a state public health department to randomly select and sample a percentage of all indivi duals in a city on a rolling basis to assess local infection rates and trends.

In this context the term *surveillance* refers to obtaining and analyzing information at the level of the group or population – without attempting to monitor or diagnose individuals – for the purpose of understanding disease incidence rates (also referred to as base-rates or priors in Bayesian analysis). This usage can be thought of as a form of screening, and not as a form of diagnostic testing. The objective of surveillance testing of SARS-CoV-2 appears to be to gain information about disease prevalence, which can be used to optimize a multiple testing strategy, and which can also be used as prior information to calculate the Bayesian posterior likelihoods of correct or incorrect results.

Pooled testing was first described by economist Robert Dorman in a 1943 article in the *Annals of Mathematical Statistics*, titled "Detection of Defective Members of Large Populations" (an atrocious title for the social sensibilities of today). The context for that publication was the United States Public Health Service and the Selective Service System, and the screening of men for Syphilis as they were being inducted into the U.S. military during WWII. In economic terms, the optimal number of samples that should be pooled together can be calculated mathematically as a function of several factors, including the incidence rate if known (or the known rate of positive test results if not), cost of testing, and expected cost savings resulting from negative analytic pools. In practical terms the cost of testing includes the social and economic costs associated with an un-contained pandemic.

In summary, an obvious advantage of multiple testing is that it can substantially reduce expenses associated with limited testing supplies such as swabs, reagents, and testing equipment, as well as the time and workload demands placed on professionals. A potential disadvantage of multiple testing is that it can constrain the types of conclusions that can be made. To the degree that it can provide adequate precision, in terms of required test sensitivity or specificity rates, or an ability to constrain FP or FN errors to required levels, multiple testing strategies can be a viable method of making maximum use of available resources, including material supplies, equipment, and human effort. As is often the cases, successful use of these strategies will depend, to some extent, upon policy administrators, field practitioners and a public that possess some knowledge or appreciation for the issues that influence scientific tests and their use.