BACKGROUND: Research in the field of proteomics to discover markers for detection of cancer has produced disappointing results, with few markers gaining US Food and Drug Administration approval, and few claims borne out when subsequently tested in rigorous studies. What is the role of better mathematical or statistical analysis in improving the situation?

CONTENT: This article examines whether a recent successful Netflix-sponsored competition using mathematical analysis to develop a prediction model for movie ratings of individual subscribers can serve to improve studies of markers in the field of proteomics. Netflix developed a database of movie preferences of individual subscribers using a longitudinal cohort research design. Groups of researchers then competed to develop better ways to analyze the data. Against this background, the strengths and weaknesses of research design are reviewed, contrasting the Netflix design with that of studies of biomarkers to detect cancer. Such biomarker studies generally have less-strong design, lower numbers of outcomes, and greater difficulty in even just measuring predictors and outcomes, so the fundamental data that will be used in mathematical analysis tend to be much weaker than in other kinds of research.

CONCLUSIONS: If the fundamental data that will be analyzed are not strong, then better analytic methods have limited use in improving the situation. Recognition of this situation is an important first step toward improving the quality of clinical research about markers to detect cancer.

© 2009 American Association for Clinical Chemistry

The Netflix Challenge
The field of proteomics research to discover markers to detect cancer has produced many claims of tests with highly accurate diagnostic discrimination, but the US Food and Drug Administration (FDA) has approved few markers, and few claims are even reproducible when subsequently tested in rigorous studies. What reasons can be offered to explain slow progress in this field, and is there a way to improve the reliability, efficiency, and overall progress of research—perhaps through the use of some kind of “best practices” to conduct research about proteomics markers for cancer?

The research challenge, briefly stated, is how to predict something unknown (e.g., the presence of cancer) based on something known (e.g., a marker that can be measured in blood). A news story about the successful application of prediction research in a nonmedical field was reported recently in The New York Times (1). A colleague wondered whether lessons from that story could apply to the field of proteomics. Addressing his question sheds light on challenges faced by the field of proteomics to detect cancer and how to deal with them.

The news story described how the movie-rental firm Netflix sponsored a contest to build a model that could outperform by 10% the model Netflix had developed to predict which movie recommendations a customer would enjoy (something unknown) from customers’ past choices and ratings of movies (something known). The success of the winning group was attributed to analytic expertise: “collaboration . . . [of] people with complementary skills and . . . different methods of problem-solving, [including] statisticians, machine learning experts, and computer engineers . . . .” An AT&T research investigator said the “contest will be looked at for years by people studying how to do predictive modeling” (1).

It is certainly fair to ask whether the approach of better computers, statistics, and problem-solving techniques can help the field of cancer markers. But first we need to understand that the Netflix approach was bigger than the mathematical competition reported in the Times. Netflix in effect conducted a research study, of

1 University of North Carolina at Chapel Hill, Chapel Hill, NC.
* Address correspondence to the author at: CB 7080, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7080. E-mail ransohof@med.unc.edu.
Received September 21, 2009; accepted November 25, 2009.
Previously published online at DOI: 10.1373/clinchem.2009.126698

2 Nonstandard abbreviations: FDA, US Food and Drug Administration; NCI, National Cancer Institute; PLCO, Prostate, Lung, Colon, Ovary study; RCT, randomized controlled clinical trial; WHI, Women’s Health Initiative; CARET, Carotene and Retinol Efficacy Trial; PRoBE, prospective-specimen-collection, retrospective-blinded-evaluation; CPTAC, Clinical Proteomic Technology Assessment for Cancer; caHUB, Cancer Human Biobank.
which mathematical analysis was just one part. In other words, the data that ended up being used in the competition was the product of a research study that gathered the right data. The Netflix study had a design—a design that may look so straightforward and commonsense that it does not need to be discussed. However, considering its structure and strengths is critical to help us understand weaknesses in current marker studies to detect cancer. Design problems solved easily in the Netflix study pose serious logistic challenges in biological and clinical research to develop diagnostic tests for cancer.

In the language of clinical research design, Netflix did a longitudinal cohort study in which customers were followed over time to gather data about movie choices and ratings. Those data could be analyzed to see whether past choices and ratings (something known) could predict future ratings (something unknown). Once these data were collected, the remaining task—and focus of the competition—was to figure out how to best analyze the data. Netflix provided identical sets of data about possible predictors and outcomes to different groups of contestants who competed to relate predictors to outcome.

**Strengths and Weaknesses of the Netflix Approach**

It almost goes without saying that the data collected for the Netflix analysis were strong: People could be followed from a time they did not already have the outcome, and possible predictors could repeatedly be measured in a straightforward and reliable way. Markers were plausibly related to outcome, whereas predictors found in current discovery research for cancer markers (2) may have no plausible relationship to outcome, perhaps explaining part of the poor record of discovery in this field. Hypothesis-free discovery may provide novel insights into biology (2, 3), but failures must be expected and weeded out aggressively. In Netflix, it also helped that each customer could have multiple outcomes, perhaps making prediction easier.

Not only were predictors and outcomes reliable and easy to measure, but the Netflix study design—a longitudinal cohort—helps avoid many kinds of bias (4–6) that plague the cross-sectional design commonly used to study diagnosis, for example when convenience samples (e.g., cases from one source, controls from another) may have a difference or signal hard-wired into specimens and thus into the comparison, producing apparent discrimination that is totally unrelated to biology or cancer (5, 7, 8). Although a detailed discussion of study design and bias is beyond the scope of this article, a brief consideration can serve to illustrate why the Netflix competition was successful and derive lessons for studying cancer markers.

The main lesson is that the mathematical analysis could be strong and produce a successful result because the fundamental data, i.e., the data on which analysis was done, were already strong. In contrast, getting strong data in a clinical study is often much more difficult.

**Using a Longitudinal Cohort Design to Study Markers for Diagnosis**

Whereas a cross-sectional design is typically used in studies of diagnosis, because the goal of a diagnostic test is to assess presence of disease at 1 point in time, diagnosis may also be assessed within a longitudinal cohort design. (Studies of prognosis or response to therapy typically use longitudinal design; they have features that differ from studies of diagnosis, even if similar principles apply.) In using a longitudinal study to assess markers of diagnosis, a nested case-control analysis has strengths that help avoid many kinds of bias (6, 9), but not all. In the Netflix study, for example, the bias of informative censoring could occur in which unhappy clients decide to go to another company and, by dropping out, remove themselves from analysis. Bias and study design are large topics, and using a longitudinal cohort design addresses only some problems. Last, problems related to design and bias are entirely separate from whether subjects chosen for study—regardless of design—will be clinically relevant or biologically relevant to answer a specific research question.

**Can a Netflix-Type Longitudinal Cohort Study Be Done to Study Cancer Diagnosis?**

Even if a longitudinal cohort study is not required to study markers of diagnosis (though it has been advocated (6)), it is fair to ask whether a longitudinal study, like that of Netflix, could be done to study cancer diagnosis, even if cumbersome and expensive. Such a study would look like this: Follow a cohort of people that do not yet have cancer. Over time, measure markers (like a serum test, something known) that might indicate presence of cancer. Frequently monitor, with some other test, for presence of cancer (something unknown). Then apply mathematical analysis to relate possible predictors and outcomes. In this approach, the study would be longitudinal, although the data would be used in a cross-sectional way to assess diagnosis near to the time blood was drawn.

Such a study has actually been done. The National Cancer Institute’s (NCI’s) PLCO study (Prostate, Lung, Colon, Ovary) is a randomized controlled clinical trial (RCT) designed to assess whether screening for 4 different kinds of cancer reduces cancer mortality.
(10, 11). A RCT is a kind of cohort study in which people are followed over time to see if they develop an outcome. A biospecimen repository included blood samples drawn repeatedly from subjects as they were followed. Some eventually developed the outcome of cancer. Among those in whom blood samples had been drawn around the time of (and before) cancer diagnosis, blood could be analyzed to learn what predictors were present when the cancer was present and so could diagnose it. Other examples of longitudinal cohort studies that have been used to study cancer markers include the Women’s Health Initiative (WHI) and the Carotene and Retinol Efficacy Trial (CARET) (12).

Data collected from a cohort can then be analyzed as a nested case-control study (a case-control study nested within a cohort) in which possible predictors in subjects with cancer can be compared to those without cancer. The nested case-control design, used for decades in clinical epidemiology, has been recently discussed for application in studies about molecular markers (4, 9). The design, recently termed the PRoBE approach (for prospective-specimen-collection, retrospective-blinded-evaluation) (6), is being applied to a study of proteomics markers to detect ovarian cancer that uses specimens from the PLCO cohort (13). The PRoBE approach, then, is less a new methodology and more a delineation of how and when the method might be used in marker research.

So if the Netflix approach of a longitudinal cohort study can be used to study molecular markers, what then is the problem? Why are we not at the same place as Netflix, where we have plenty of good data and just need to figure out better methods of analysis?

The answer is instructive—if frustrating for now—and will, over the long run, help us improve the reliability and efficiency of how we study molecular markers. The details of how to improve go beyond what can be addressed here and are considered elsewhere for molecular markers (4, 9, 14, 15) and for markers in general (16–19). The main theme or lesson here is that important methodological design features that are easy to implement in the Netflix study are much harder in a clinical or biological study.

How Feasible Is the Design of a Longitudinal Cohort?

Getting good data for a clinical study is more difficult—by orders of magnitude—than for the Netflix study. Studies like PLCO and WHI may cost hundreds of millions of dollars and take decades to conduct. Even after substantial investment of time and effort, the yield of relevant data may be relatively small. After following more than 140 000 people for >10 years, the cancers occurring in PLCO number in the low hundreds, and the numbers of people with blood samples drawn near to the time of (and before) diagnosis are lower. Low numbers have practical consequences. To make a predictive system, the general rule of thumb is that you need 5–10 outcomes (such as cancer) for every candidate predictor. If you are looking for 10 possible predictors, that translates to a need for 50–100 outcomes—a huge number in clinical research terms. Finally, another limited quantity is the amount of blood or serum that can be obtained from each subject; there are not enough milliliters of blood or serum to be given to provide identical specimen sets to more than a small number of laboratories to develop or test some diagnostic assay. In contrast, the number of outcomes in the Netflix study was enormous: 100 million movie ratings (20). And the number of identical data sets that could be distributed for analysis to research groups is virtually limitless; after all it’s just electrons. More than 40 000 groups competed simultaneously in the Netflix contest (21).

If a Longitudinal Cohort Cannot Be Done, How Can Bias Be Avoided?

Although the longitudinal cohort study design avoids many biases, such a study is expensive and difficult to do. Other designs risk having important inherent biases that may invalidate the comparison (cancer vs not) and introduce spurious signal. One colleague recently asked, after noting that bias is increasingly discussed, “Since we have learned a lot about bias in molecular marker studies in the last few years, doesn’t that mean that we know how to avoid it, perhaps using some kind of ‘best practices’ in clinical research or specimen collection?” The answer is that what we have learned in the last few years is not about new kinds of biases that, now identified, are solvable. In the last few years we have simply recognized that the old biases that cause serious problems in other kinds of observational epidemiology research also cause the same problems in biomarker research. Recognizing that they exist does not make them easier to solve. Recognition is an important step, to be sure. But the topic of bias—and how to avoid it—is the stuff of courses, textbooks, and careers in clinical research. Rules of evidence and principles can be usefully applied, but application must be done case by case in a way that considers specifics of the biology and technology being assessed (4, 5). No list of best practices or set of guidelines is available that can be meaningfully applied by someone who is not deeply trained in the area, and problems may be substantial even for those with deep training.

Other Challenges: Identifying Predictors and Outcomes

In marker research, and proteomics in particular, how well can possible predictors be measured? In the Netflix
study, predictors are easy to measure—past choices and ratings; common sense tells us roughly what kinds of predictors to look for. (In a new iteration of the contest, Netflix will assess other potential predictors such as demographics and behavioral data (20).) Developing scales of measurement is not difficult, and different people trying to measure the same thing would likely come up with the same answer. In the field of proteomics to detect cancer, none of these things is straightforward. We may even be uncertain about what we are looking for: Where in the proteome is the information that we are trying to find? Is it in big proteins or small ones? Antigens or antibodies? High abundance or low? Normal proteins, or protein variants unique to cancer? And even if we knew the answers to those questions, we do not have agreed-upon and reproducible methods for measurement; there are few reference standards that help different laboratories assess whether what they think they have measured is similar or different from what some other laboratory has measured. These limitations in the field of serum proteomics are only now starting to be addressed (22).

Even measuring an outcome like cancer (and in particular early and potentially curable cancer) in a clinical study is not straightforward. Special follow-up systems need to be put in place, perhaps involving a cumbersome or invasive imaging test followed by workup. Without such a system, only late-stage cancers would be discovered. To reliably measure whether heart attacks, including “silent” heart attacks, had occurred in the Framingham Study longitudinal cohort, elaborate systems had to be instituted to measure the outcome (23, 24).

**Lessons, and How to Apply Them**

Discussing lessons and implications risks being frustrating, but discussion is obviously necessary to address problems.

The first lesson is to appreciate that the main problem—or at least a critical rate-limiting problem—is not mathematical strength or technological strength. The problem is clinical research design that, if faulty, produces specimens that, when analyzed by technology and mathematics, produce misleading results. If the fundamental data have bias hard-wired into them, erroneous conclusions become the weak foundation for subsequent work and wasted effort (4, 5, 7).

Second, there are ways to deal with bias, but they are not easy. There certainly is no checklist or set of best practices that can be handed to biochemists, bench researchers, or clinicians for implementation. At best, there are general guidelines and rules of thumb for design (14, 15, 17, 25), but these may be challenging to apply even for those with deep background in clinical research methods. The reality is that the entire field of methods for observational research to study tests and markers is not well developed (4). As an example of the magnitude of work that may be required, methods of the case-control study took decades to evolve, as biases came to be understood by trial and error and difficult experience. After dozens of biases were identified and studied (26), case-control studies are now better understood, are better designed, and are considered to be more reliable than in the past. The field of research design to study markers is not starting from scratch, to be sure, but there is much work to be done to identify biases and determine how to deal with them, particularly when an “ideal” study (like a longitudinal cohort) cannot be done.

Third, some recent developments hold promise. Longitudinal cohort studies—some existing and some that may be created in the future—may be available for both discovery and validation. The need for strongly unbiased specimens in discovery is increasingly discussed in NCI’s EDRN (Early Detection Research Network) and CPTAC (Clinical Proteomic Technology Assessment for Cancer) and in commentaries about methodology (4, 6, 27). NCI’s PLCO has made available specimens from its biorepository for both validation and discovery (13). Perhaps other major specimen-collection efforts, like NCI’s caHUB (Cancer Human Biobank) now being planned, may see this national need as an opportunity.

Finally, the Netflix story may provide an important big-picture lesson about how to conduct certain kinds of translational or interdisciplinary research. By the time the statisticians received Netflix data to analyze, the rest of the study had been done—and done well—by others; importantly, the fundamental data received by the competing teams were sound. The statisticians’ task, then, was circumscribed and clear. Perhaps a similar “divide and conquer” approach should be used more deliberately in research about markers for cancer (7). Rather than trying to make bench researchers or technologists (who are now leading much marker research) into clinical epidemiologists who can do both kinds of research, maybe it would be smarter to try to separate—rather than combine—the 2 kinds of research. Let the clinical researchers and clinical epidemiologists collect specimens, with proper attention to bias, clinical use, and so on; then those specimens can be handed off to the technologists and statisticians. Some cross-talk or discussion between the disciplines of clinical research science and bench science will of course be needed. But it may be more productive in the long run to clearly separate disciplines and let each do what it does best (7). One kind of product might be, like Netflix, a large repository of specimens that can be used by multiple groups simultaneously for both dis-
covery and validation (4). Others may disagree with the strategy of separation and say that technologists should be further educated, and there may be several ways to skin the cat. In the meantime, however, we need to consider that the arrangement used over the last 10 years has not been particularly successful. At the end of the day, the data used in analysis must be basically sound. If those fundamental data are strong, the possibilities for reliable exploration will be limitless.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

References