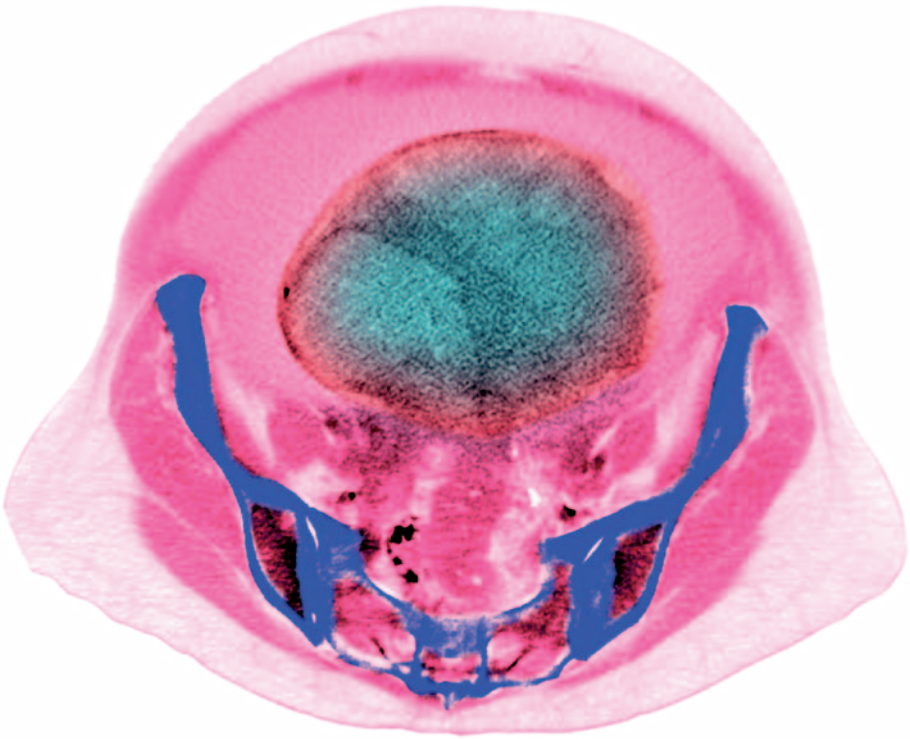
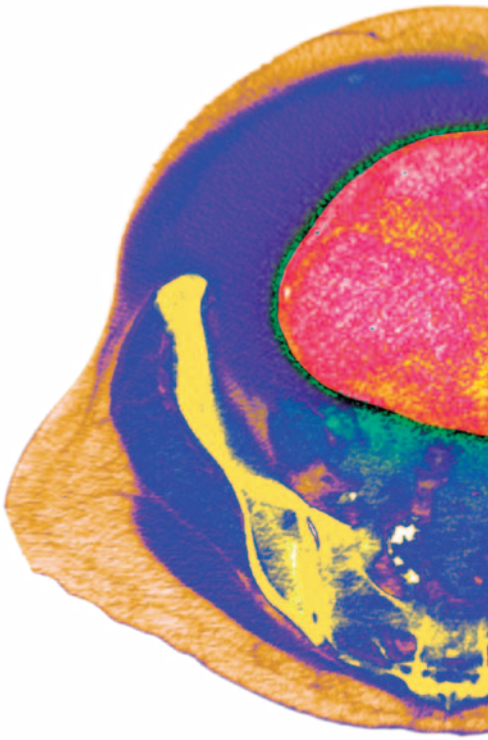




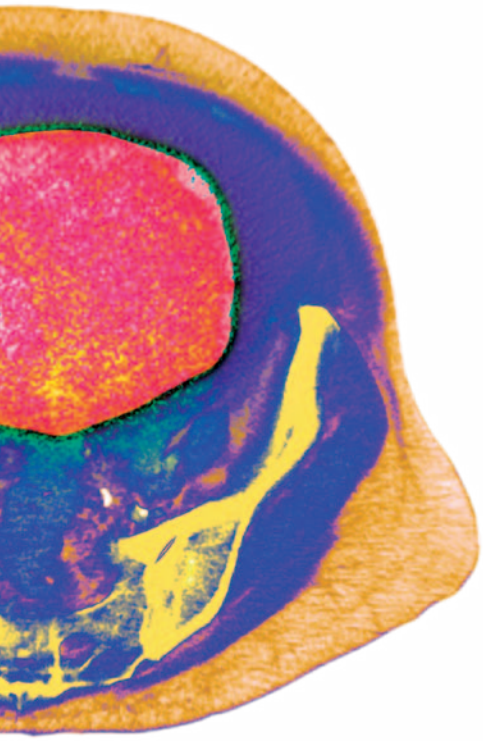
MISSING THE MARK

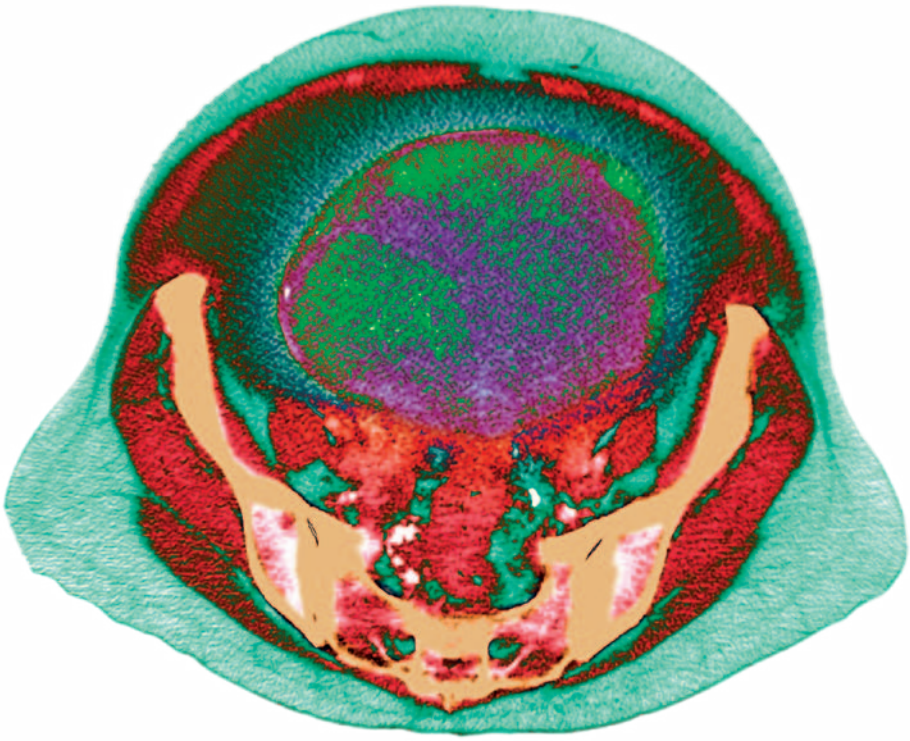
Why is it so hard to find a test to predict cancer?

BY LIZZIE BUCHEN

On 3 March, two studies appeared online that offered 19 pages of gloomy reading for anyone interested in cancer. They focused on biological molecules, or biomarkers, the presence of which in the blood might be used to detect the earliest glimmers of ovarian cancer — a disease not normally discovered until it has destroyed the ovaries and rotted other parts of the body. The researchers, coordinated by the Early Detection Research Network (EDRN) of the US National Cancer Institute (NCI), had assembled 35 protein biomarkers, including 5 panels of proteins, that had looked the most promising in early studies. They had carried out rigorous testing — screening blood samples from more than 1,000 women — to ask whether these seemingly breakthrough biomarkers were better at identifying women with early ovarian cancer than the one flawed biomarker that had been in use for almost 30 years, CA-125. None of them was[1,2]. "CA-125 remains the 'best of a bad lot'," read an accompanying perspective article[3]. "The new candidates have fallen short of expectations."

Tied in last place for its poor performance among the biomarker panels was one identified by Gil Mor, a cancer biologist at Yale University in New Haven, Connecticut. Mor's six-protein panel detected ovarian cancer in only 34% of the women who were diagnosed with the disease within a year. (CA-125, by contrast, detected 63%.) Mor's panel already had a tortured history. A primary research paper behind it had been criticized by other scientists for allegedly using inappropriate statistical calculations and for optimistically concluding that the test would help women before rigorous follow-up studies proved that it could. Yet for four months in 2008, the test was sold to patients by Laboratory Corporation of America (LabCorp) in Burlington, North Carolina, the company that licensed the panel from Yale. LabCorp had marketed the test under the name OvaSure until the US Food and Drug Administration (FDA) intervened and the company pulled it from the market. The panel offered "invaluable object lessons"





IMAGES BY DU CANE MEDICAL IMAGING LTD/SPL

Biomarkers might help to detect ovarian tumours (large round masses) early in the disease.

for bringing a test prematurely to the clinic, wrote the authors of the perspective article.

Similar lessons can be found in the stories behind many cancer biomarkers that have sputtered and failed on their way to the clinic. Those tests that are in clinical use — including prostate-specific antigen (PSA) for prostate cancer, mammogram-detected masses for breast cancer and CA-125 — fail to detect all cancers and sometimes 'detect' ones that aren't there. Genomics, proteomics and other such technologies promised to help by finding combinations of markers that are more powerful and cancer-specific than individual ones, but that promise has not been realized. Researchers using such technologies have published studies on thousands of panels, suggesting that they can detect early-stage disease, guide patient treatment and monitor recurrence. But only a tiny number of such tests have reached the clinic — and none for the early detection of cancer, the biggest clinical challenge of all. "Much biomarker research has been done very badly for decades," believes Lisa McShane, a biostatistician at the NCI in Rockville, Maryland. "Even when it was single markers. Now, as we're moving up to multiple markers, all our bad habits are coming back to bite us in a big way."

These habits have been thrown into the spotlight by the EDRN's study, one of the largest and most systematic validation studies of biomarkers so far. It came just months after a high-profile decision at Duke University in Durham, North Carolina, to suspend clinical trials of a genomics-based biomarker panel designed to direct chemotherapy in patients with breast cancer. A number of scientists had raised concerns about the Duke group's data and analysis, and the trial was stopped after allegations came to light that the lead researcher, geneticist Anil Potti, had made false claims on his CV. Last September, the Institute of Medicine (IOM), part of the US National Academies, assembled a committee to discuss lessons for developing tests based on 'omics' technologies and bringing them to the clinic. "Why don't we have assays out there, with this enormous promise?" Dan Hayes, a breast-cancer researcher at the University of Michigan in Ann Arbor asked researchers at the first IOM committee meeting in December 2010. "It's either because these things just don't work, or because we've used sloppy science to test them."

It is too early to say whether either of these is true: the field is still young, and faces many challenges. It has drawn in many cancer biologists who are excited by the potential to translate their work to the clinic — but they sometimes lack the expertise or resources needed to pursue translational or clinical work. "A lot of novices came in. They get in without realizing that the problem may be more complex than it appears," says Eleftherios Diamandis, a clinical biochemist at the University of Toronto in Canada. And although most experts agree that potential biomarkers for early cancer detection should be validated on samples taken before diagnosis — the stage at which the test would be used in the clinic — that is a step that few groups attempt and no biomarker for ovarian cancer has passed, as the EDRN study made clear. "Sometimes the glamour of the technology or the sheer volume of omics data seem to make investigators forget basic scientific principles," said McShane at the IOM meeting. Mor agrees that the field has faced problems, and that it is important for markers to go through a careful process of design and validation, as he tried to do.

"There's been an enormous amount of hype and promise," sums up David Ransohoff, a cancer epidemiologist at the University of North Carolina in Chapel Hill. "But after 10 or 15 years of intense work in these fields, there's simply not a lot to show for it. It's important for the whole field to step back and look at what is wrong."

NATURE.COM
A possible way forward for cancer biomarkers:
go.nature.com/icwtue

MAKING A DIFFERENCE

Mor began his career in Israel, where he trained as a clinician at the Hebrew University of Jerusalem. But an experience in the final years of his oncology residency compelled him to change course. A young woman arrived at the hospital with ovarian cancer, a disease that kills some 140,000 women worldwide each year. The oncology team removed the woman's ovaries and put her through several rounds of chemotherapy, which seemed to be successful. But 18 months later, she was back, her body riddled with tumours, and she soon died. "Chemotherapy didn't do anything for her,"

Mor recalls. "She was 29. She was a beautiful girl. An impressive girl. A medical student. And I never understood what happened to her."

Mor decided to leave medicine, which had been unable to save her, for research, which one day might. He earned a PhD studying ovarian cancer at the Weizmann Institute of Science in Rehovot, Israel, before moving to Yale in 1997. He went on to start a programme called Discovery to Cure, aiming to speed cancer research to the clinic. The group began to build a bank of blood and tissue samples, including some from a Yale clinic for women with a high risk of ovarian cancer owing to a family history of the disease. "There was a lot of excitement around that time for finding proteins specific to cancer," says Mor.

"AS WE'RE MOVING UP TO MULTIPLE MARKERS, ALL OUR BAD HABITS ARE COMING BACK TO BITE US IN A BIG WAY."

In 2003, David Ward, then a geneticist at Yale, contacted Mor. Ward had co-founded Molecular Staging, a company in New Haven that had developed a 'high-throughput' technique for quantifying multiple proteins in the blood using arrays of antibodies[4]. He asked whether he could use Mor's samples to search for markers of early ovarian cancer.

Mor had never been involved with biomarker research — "I do biology of cancer, not biomarker development," he says — but he signed up, intrigued by the clinical potential of the technology. Ward had scoured the literature for proteins that had been associated with ovarian-cancer growth and malignancy, and had come up with 169 candidates. Using the protein-quantification technique, Ward's company screened blood samples in Mor's tissue bank that came from two groups: women with newly diagnosed ovarian cancer who had been enrolled in Yale's high-risk clinic, and women who had come to the hospital for routine gynaecological exams. Using additional cancer-patient samples, they whittled the list down to four proteins: leptin, prolactin, osteopontin and insulin-like growth factor II.

Mor worked to develop an algorithm that could automatically classify women as having cancer or not, depending on levels of these four proteins. When the team ran a new set of blood samples through the algorithm, they got astounding results. The test showed a sensitivity of 95% (meaning it correctly detected 95% of the ovarian-cancer cases) and a specificity of 95% (it erroneously classified only 5% of healthy people as having cancer). "I was delighted," says Mor. On equivalent samples, CA-125 tests typically have a sensitivity of 70–80% and a specificity of around 95%. In May 2005, the findings were published in the *Proceedings of the National Academy of Sciences* (*PNAS*), with Ward as a contributing author[5].

Before publication, Mor helped the Yale Office of Cooperative Research to prepare a patent application. "A lot of companies expressed interest in licensing the panel," says John Puziss, director of technology licensing at Yale. LabCorp licensed the test in 2006, as did Millipore, a biomanufacturing company based in Billerica, Massachusetts. (Mor says that the royalties he and his co-inventors received "were not a significant amount".)

The test's promising results had also caught the attention of researchers in the EDRN, who were just putting together their validation study. Up to that point, most biomarkers for detecting early ovarian cancer had only been shown to distinguish patients with diagnosed cancer from healthy controls, but they are intended to detect the disease in women whose cancer is just budding, before symptoms develop. What the field needed was a 'prospective' study, run on blood samples from apparently healthy women, to see whether the biomarkers could pinpoint those who would later be diagnosed with ovarian cancer. Such samples, from large numbers of women who are tracked over months or years, are extremely difficult to come by.

PROBLEM DETECTION

The EDRN found what was needed in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, sponsored and run by the NCI. Between 1992 and 2001, the trial had been collecting blood at regular intervals from 155,000 women and men, and screening them for cancer. By June 2006, 118 of the women had developed ovarian or closely related cancers, and the EDRN researchers were now in a position to use them to evaluate the most promising biomarkers for early detection. Ziding Feng, a biostatistician at the Fred Hutchinson Cancer Research Center (FHCRC) in Seattle, Washington, and coordinator of the EDRN, visited Mor to discuss whether his panel of four proteins could be included in the study.

Mor was already in the process of refining the panel: he had more patient samples, and wanted to add more markers, including CA-125 and the protein macrophage migration inhibitory factor, to make the test more sensitive to cancer. LabCorp had been running his new samples on assay kits manufactured by Millipore. (Ward, meanwhile, had moved to the Nevada Cancer Institute in Las Vegas, and was not involved in data collection or analysis.)

When Mor showed Feng how he was analysing his recent data, Feng was troubled. Mor asked him to go through the new results himself, and Feng agreed to collaborate. "I do not do statistics," says Mor. "That is not my field." The researchers also added the six-protein panel to the EDRN's validation study.

Feng and Gary Longton, another statistician at the FHCRC, developed their own classification algorithms, and found that Mor's test had a sensitivity of 95% and specificity of 99%. They also calculated the positive predictive value (PPV) of the test — the proportion of patients who the test would diagnose with the disease and do in fact have it. A high PPV means that few people will be misdiagnosed, which is crucial when screening healthy people.

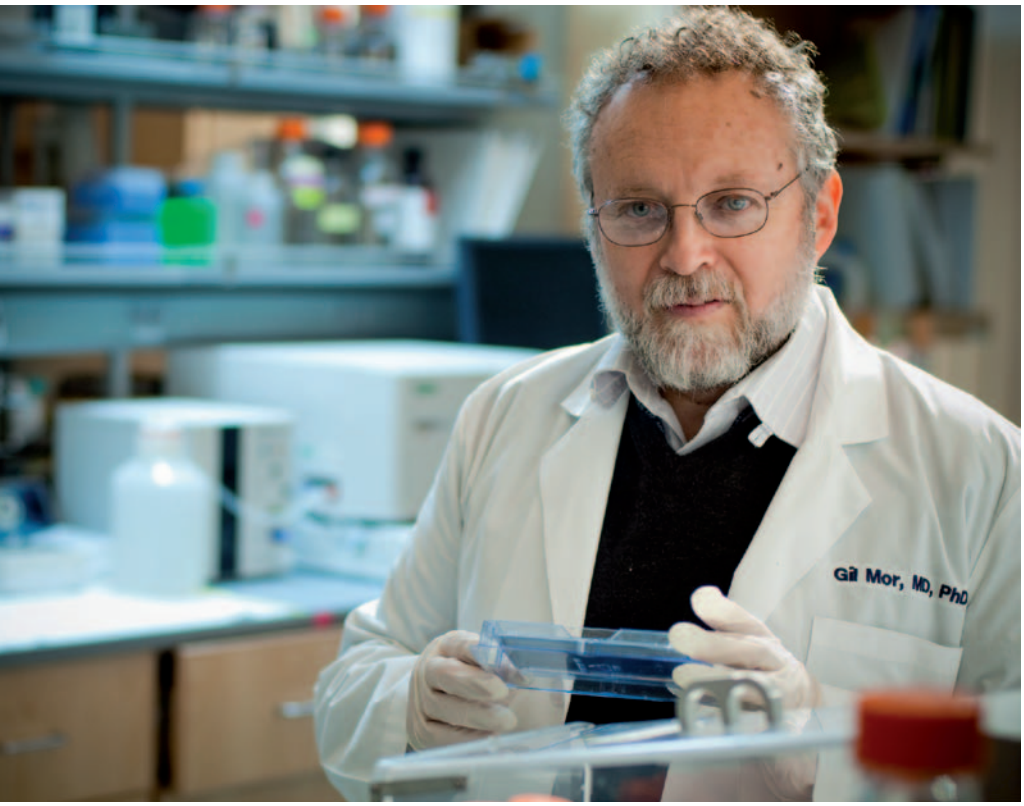
Feng and Longton calculated the PPV at 6.5%, too low for the test to be of much use for screening. But separately, Mor was working with a different figure, of 99.3%. The huge disparity between the two values stemmed from the way that they calculated the figure and factored in the prevalence of ovarian cancer — an important variable in calculating the PPV. Following convention, Feng and Longton calculated the PPV using the accepted prevalence in postmenopausal women, 1 in 2,500 (0.04%). But Mor's figure was calculated solely from the study population, in which the prevalence was 46%. "We calculated the PPV based on the population in the study, because we always intended the test for the high-risk population," says Mor. "If you want to bring the test to the clinic, it has to be calculated based on the population you're going to study," he says, noting that other research studies work out the PPV for the study population in this way.

It's a common mistake, believes McShane, who — like other statisticians — disagrees with Mor's logic. "I see that a lot, but it is nowhere near the correct thing to do," she says. Even in high-risk populations — women who are screened every year because of their family history or because they have tested positive for mutations in tumour-suppressor genes *BRCA1* or *BRCA2* — the prevalence is around 0.5%, far below the 46% in Mor's study population. Similar battles over the correct use of statistics litter the cancer-biomarker field, said researchers at the IOM meeting last year. "It's the type of thing where non-statisticians think statisticians are being uptight about something that's not going to matter anyway," says McShane.

"IT'S IMPORTANT FOR THE WHOLE FIELD TO STEP BACK AND LOOK AT WHAT IS WRONG."

Mor prepared a paper reporting the latest work. But when Feng and Longton saw the page proofs, they noticed that the PPV value was reported as 99.3%. They asked Mor to change it to the 6.5% that they had calculated, and to correct a few other typographical errors in the tables. "He agreed, so we signed off," recalls Feng. But there was a miscommunication: Mor thought that Feng had agreed to the use of the high PPV, and that everyone approved of the final manuscript.

The paper was published online in *Clinical Cancer Research*[6] in February 2008, and to



Gil Mor is testing whether a panel of six proteins can detect ovarian cancer in women at high risk.

S. OGDEN

Feng's shock it reported the high PPV. "You can imagine how upset I was when I saw it in the paper," says Feng.

Feng called Mor. "I told him, those are errors, we told you those are not correct." Feng also contacted the journal, the editor of which asked Mor to submit a correction to fix the PPV and the other typos. Mor agreed, adding the lower PPV as a footnote to the table and in a written correction.

A few weeks later, Feng received an e-mail with unwelcome news from a colleague: LabCorp was preparing to market the panel, and was "hopeful that this test will be available to women by the end of the year".

"I was shocked," says Feng. "I had no idea this was coming." He thought that the markers should be validated further before they went to the clinic. In March 2008, Feng and Mor saw each other at a meeting in Washington DC. "I told him, face to face, you cannot do this," says Feng. "You have to wait until after the PLCO validation. What you have done is early discovery. If validation does not support your earlier claim, you're making a significant error." Mor does not recall this encounter, but says that Feng's "role was to analyse the data, not to make judgements of a company decision".

Now, Mor says that if he were preparing the paper again, he would include both the low and high values for the PPV. And he vacillates about whether LabCorp's decision to offer the test to women before it had undergone more validation studies was the right thing to do. He says he thought that clinical use of the test might be a good way to do further validation. "It's very difficult to do that on large numbers of patients," he says. "It's extremely expensive. The only way to do the study is if LabCorp started distributing the test and enrolling patients." Mor notes that many tests, such as mammography, have been offered to patients as an aid to diagnosis even while data on the test are being collected. "Was it the right time? I don't know," he says.

CRITICAL BACKLASH

On 23 June 2008, LabCorp announced the availability of the OvaSure test, for between US$220 and $240. The press release said that it was being offered to women with a high risk of the disease, and quoted Mor as saying he was "pleased that this test is available to help physicians detect and treat ovarian cancer in its earliest stages".

Excited chatter about the test spread through patient forums and support groups, but it was soon countered by cautionary tales. Jean McKibben, an ovarian-cancer survivor, rushed to take OvaSure on the first day it was available, and her results showed a 0.00 chance of cancer. A week later, scans showed that her cancer was back. She was crushed. "I wanted this to work so badly," she wrote on a discussion board.

One week after LabCorp's announcement, the Society of Gynecologic Oncologists in Chicago, Illinois, released a statement expressing concern about OvaSure, saying that "additional research is needed to validate the test's effectiveness". The paper in *Clinical Cancer Research* was also circulating at the Canary Foundation, a non-profit organization based in Palo Alto, California, that funds research on early cancer detection. Scientists there found other reasons for concern. One member, Nicole Urban, head of the Gynecologic Cancer Research Program at the FHCRC, had found that levels of prolactin, one of the proteins in the panel, are highly sensitive to stress — something very likely to affect women entering the clinic with symptoms of ovarian cancer[7]. After controlling for that, she says, "prolactin gave no signal at all for malignancy. It was useless." Others pointed out that the high specificity and sensitivity figures reported in the paper's conclusions, and trumpeted in Yale and OvaSure press releases, were not present in any of the tables or figures. And they bristled at the positive tone of the discussion, which stated that the test "will enhance the potential of treating ovarian cancer in its early stages and therefore, increases the successful treatment of the disease".

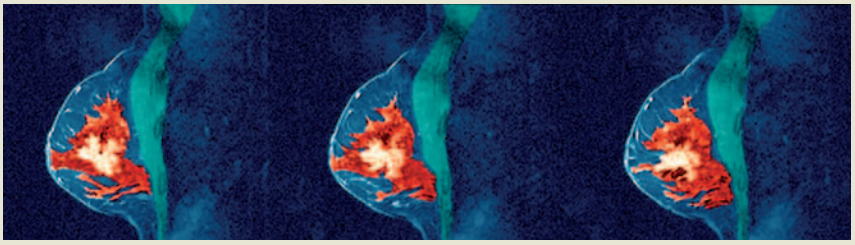
"There were a lot of uncertainties, and evidence of biases," says Martin McIntosh, who researches markers for early-stage ovarian cancer at the FHCRC, and is a member of the Canary group, "But the narrative only highlighted the best-performing analysis. It didn't mention caveats." Members of the Canary group wrote a letter to *Clinical Cancer Research*, describing some of their complaints. Meanwhile, Feng agreed to co-author a second letter, criticizing the paper even though he was a co-author.

The fuss was already reaching the FDA, which on 7 August 2008 sent a letter to LabCorp saying that the test "has not received adequate clinical validation, and may harm the public health". A second letter, sent by the FDA on 29 September 2008, alleged that LabCorp did not have the necessary marketing clearance or approval for the test from the FDA. LabCorp replied to the FDA on 20 October, disagreeing with the agency's assertions, but agreed to pull OvaSure from the market. It did so on 24 October 2008, just one day after *Clinical Cancer Research* published the critical letters from the Canary Foundation and Feng, as well as a third from the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia[8–10]. (Millipore continues to market the biomarker panel for use in research, not by patients.)

Mor was surprised by all three letters. In his published response[11], he disputed some of the criticisms and wrote that any concerns about commercialization should be taken up with LabCorp. Stephen Anderson, vice-president of investor relations at LabCorp, says that OvaSure was not marketed as a test for detecting cancer recurrence, which was how some patients used it. He says that LabCorp "continues to believe OvaSure offers a

CASE STUDY

The gene collection that could



ZEPHYR/SPL

Genomic signatures measured in tissue samples can help to classify breast-cancer tumours.

When asked to name a successful cancer test that is based on multiple genes or proteins, many researchers point to Oncotype DX. The panel, which tests surgically removed breast tumours for the expression level of 21 genes to predict the likelihood of the cancer recurring, was developed by Genomic Health in Redwood City, California, and has been marketed since January 2004. It is used by roughly half the patients in the United States who have the most common type of breast cancer.

"A lot of biomarker research starts with getting some signature and then figuring out how to use it," says Richard Simon, a biostatistician at the US National Cancer Institute in Rockville, Maryland. But Genomic Health researchers took a different route, sitting down in 2000 with a group of oncologists and patient advocates and asking what question in cancer treatment they should address. Two years later, they nailed it down. At present, the majority of women with the most common type of early breast cancer undergo chemotherapy after surgery, but only 15% are likely to have a recurrence. Was it possible to identify the crucial 15%, and spare the others from chemotherapy? The team set out to find a gene signature that could do the job. "The number one key to their success was starting with a well-defined, clinically relevant question," says Simon.

The researchers brought in Michael Walker, a statistician then based in Sunnyvale, California, to help design the studies from the outset. Walker says that it rarely works this way, and that often the statistician is only brought in after the data have been collected. By that time, biases and confounding factors may be hard-wired into the data. The team also put a high priority on using the right tissue samples in their initial studies. They decided early on that they wanted to use tumour tissue that had been fixed in formalin and embedded in paraffin — the way it is prepared by the pathology lab after a tumour is removed — in the clinic and in all clinical trials.

Because of this, the team was able to validate the panel using samples that had already been collected in large clinical trials, rather than having to collect samples afresh.

"It's a poster child for one way to do clinical research," says David Ransohoff, a cancer epidemiologist at the University of North Carolina in Chapel Hill. But the test is hardly perfect. In January 2008, a group commissioned by the Centers for Disease Control and Prevention in Atlanta, Georgia, evaluated Oncotype DX. It found that the test results were reproducible and did well at predicting recurrence, but it was unclear whether the test was better than established risk factors, such as age, or standard molecular features of the tumour[13]. Results from a large, independent validation study, called TAILORx, are expected in 2015. **L.B.**

valuable tool for ovarian-cancer detection in conjunction with other diagnostic techniques", and that the assay is still in development. The company would not provide further comment.

DOUBTS AND LESSONS

Since then, Mor has worked hard to validate his panel. He and Ward have completed a study on a much larger set of samples including many from women diagnosed in the earliest stages of ovarian cancer[12], and in which LabCorp again ran the assays. The test still performed well at distinguishing the patients from the healthy controls. Mor says he is puzzled by the PLCO trial results, and he hopes that further analysis of the trial data will help to explain why his biomarkers performed so poorly. He continues to express confidence in his panel, saying that the test could be most useful in high-risk populations, and when used regularly — every two to three months — to monitor rising and falling levels of the biomarkers. But the whole experience has made him reluctant to pursue biomarker work much further. "I'm focusing on understanding cancer stem cells," he says.

Others say that's just as well. The panel's poor performance in the PLCO study makes critics question its usefulness in any group, even a high-risk one. McIntosh says that the PLCO study's damning conclusions should serve as a wake-up call. "The entire field has to cope with this," he says — including him, given that the most promising biomarkers discovered by his institution also failed to improve on CA-125 in the trial. "It's hugely disappointing."

The IOM committee, which is expected to release its results sometime in 2012, may help to find a way forward. At a meeting later this month, the members plan to draw lessons from the biomarker failures, as well as from the few success stories (see 'The gene collection that could'). One of the most urgent lessons is the need to help researchers validate their biomarkers on appropriate samples before they reach the clinic. Feng says that the EDRN has been collecting its own high-quality tissue reference sets for ovarian, breast, lung, colon, liver and prostate cancers, from people who aren't yet showing symptoms and those in all stages of the disease. Investigators can apply to test their biomarkers on blinded tissue samples.

Until this type of testing becomes commonplace, there is no way of excluding the possibility that, as Hayes suggested at the IOM meeting, "these things just don't work" — particularly when it comes to picking up cancer early on.

"People keep talking about early-detection biomarkers as if they are a fact, and we only need to find them," says McIntosh, "when in reality their existence is a hypothesis that needs to be tested." ■ SEE OUTLOOK P.450

Lizzie Buchen *is a freelance writer in San Francisco, California.*

1. Cramer, D. W. *et al. Cancer Prev. Res.* **4,** 365–374 (2011).
2. Zhu, C. S. *et al. Cancer Prev. Res.* **4,** 375–383 (2011).
3. Mai, P. L., Wentzensen, N. & Greene, M. H. *Cancer Prev. Res.* **4,** 303–306 (2011).
4. Schweitzer, B. *et al. Proc. Natl Acad. Sci. USA* **97,** 10113–10119 (2000).
5. Mor, G. *et al. Proc. Natl Acad. Sci. USA* **102,** 7677–7682 (2005).
6. Visintin, I. *et al. Clin. Cancer Res.* **14,** 1065–1072 (2008).
7. Thorpe, J. D. *et al. PLoS ONE* **2,** e1281 (2007).
8. Greene, M. H., Feng, Z. & Gail, M. H. *Clin. Cancer Res.* **14,** 7574 (2008).
9. McIntosh, M. *et al. Clin. Cancer Res.* **14,** 7574 (2008).
10. Coates, R. J., Kolor, K., Stewart, S. L. & Richardson, L. C. *Clin. Cancer Res.* **14,** 7575–7576 (2008).
11. Mor, G., Schwartz, P. E. & Yu, H. *Clin. Cancer Res.* **14,** 7577–7579 (2008).
12. Mor, G., Symanowski, J., Visintin, I., Birrer, M. & Ward, D. *Proc. Am. Assoc. Cancer Res.* LB-224 (AACR, 2009).
13. Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group *Gen. Med.* **11,** 66–73 (2009).