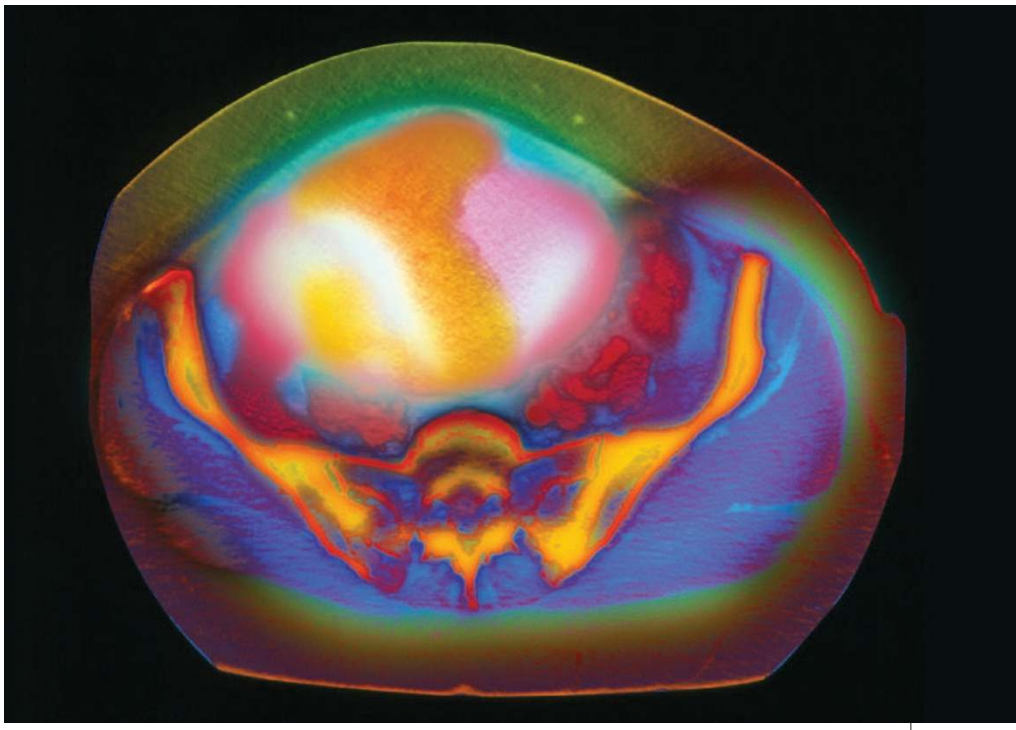
Running before we can walk?

Two years ago, a new proteomic test was heralded as the future of cancer diagnostics. But since then, doubts about its effectiveness have begun to grow. Erika Check reports.



SPL

On target: can proteins in the blood reveal ovarian tumours (pink/yellow) before they reach this stage?

Seldom does a single piece of research prompt the US Congress to pass a resolution urging continued funding to drive a new diagnostic test towards the clinic. But that's what happened in 2002, when *The Lancet* published a paper[1] claiming a breakthrough in the diagnosis of ovarian cancer.

The paper described the use of mass spectrometry to analyse the pattern of proteins present in samples of blood serum. On the basis of these patterns, the test detected all the patients with ovarian cancers in a set of 50 samples, and falsely identified just three healthy patients as suffering from the disease from a total of 66 control samples.

Most encouragingly, the technique seemed to work well on patients with early-stage cancer — offering the prospect of earlier diagnosis, which improves the chances of successful treatment. The best current blood test, which relies on the detection of a single protein called CA125, misses at least half of patients in the earliest stages of the disease, and gives a high rate of false positives.

The researchers, led by Lance Liotta and Emanuel Petricoin, who co-direct a proteomics programme run by the National Cancer Institute and the Food and Drug Administration, based in Bethesda, Maryland, won immediate acclaim. In addition to the congressional resolution urging further funding for their research, the consumer magazine *Health* named the test one of the top ten medical advances of the year.

Commercial rights to develop the test were quickly licensed from the US government to a company called Correlogic Systems, also based in Bethesda, whose scientists collaborated with Liotta and Petricoin on the *Lancet* paper. In November 2002, Correlogic granted licences to two larger firms, Quest Diagnostics and the Laboratory Corporation of America, which are now hoping to market the test under the brand name OvaCheck.

But those plans could be thrown off track by reanalyses of Liotta and Petricoin's data by independent groups, which have raised serious doubts about OvaCheck's reliability.

These questions prompted the Society of Gynecologic Oncologists to review all of the published work about OvaCheck. On 7 February, the society declared that "more research is needed to validate the test's effectiveness before offering it to the public".

Early warning

Critics warn that the episode illustrates the dangers of moving rapidly to the clinic with immature technologies such as those of proteomics. They say that scientists and regulators need to develop standards to ensure that such tests really work before they hit the market, because early detection is not without risks. Women who get false positive results may undergo unnecessary surgery, and those who get false negatives may forgo further screening. "Early detection is not a benign undertaking," says Martin McIntosh, who runs a cancer-detection effort at the Fred Hutchinson Cancer Research Center in Seattle, Washington.

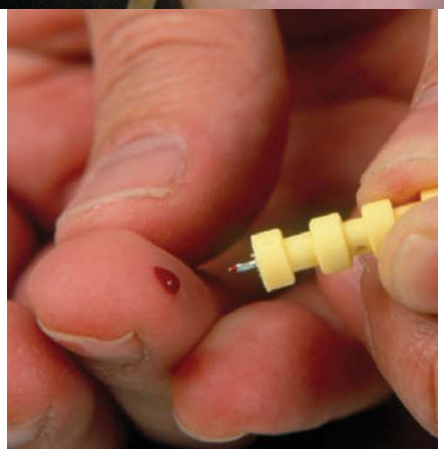
> "Whether or not OvaCheck works, we will learn from this experience what rules of evidence we might apply in the future to find useful results more efficiently."

The first criticisms of OvaCheck hit the public domain in June 2003, when two biostatisticians at the University of Maryland in Baltimore, James Sorace and Min Zhan, published a paper in the online journal *BMC Bioinformatics*[2]. They had reanalysed a data set that Liotta and Petricoin's team posted online in August 2002. Sorace and Zhan similarly found numerous differences in the protein patterns that discriminated between the cancer patients and the healthy controls. The trouble, according to Sorace and Zahn, was that these looked more like experimental artefacts than real biological differences.

The proteomics test relies on using gravity and electric fields to separate the proteins in a given sample. Each protein is then given a number that represents the ratio of its charge and mass — called its m/z value. The test identifies patterns in these numbers to give a diagnosis. Sorace and Zahn were concerned because the most marked differences between the cancer patients and unaffected controls occurred for proteins with m/z values of less than 500. Many spectrometry experts consider m/z values below 2,000 to be suspect, because they tend to include values generated by experimental artefacts or measurement error. To Sorace and Zahn, this suggested that the data for cancer patients and controls had been collected under such different conditions that it was impossible to reliably identify true biological differences.

Further questions emerged when biostatisticians led by Keith Baggerly of the MD Anderson Cancer Center in Houston, Texas, reanalysed the data in the *Lancet* paper, plus two further data sets posted online. "Based on these data, we really can't tell if it's possible to use proteomics to separate normals from cancers," Baggerly says.

K. LAMBERT/AP



SATURN STILLS/SPL

Emanuel Petricoin (above) holds a protein pattern generated by the blood test he believes can reliably diagnose cancer.

Baggerly's first concern was that the values posted on the proteomics programme's website had already been processed in a way that made it impossible to reconstruct the raw data. But more specific concerns arose when his team analysed the overall characteristics of the data, looking at how closely the peaks for each of the proteins in the samples matched one another. Each of the data sets was divided into three groups: cancer patients, unaffected controls and patients with benign tumours. In the first data set, the pattern of proteins detected in the benign-tumour group seemed very different from the patterns detected in both the cancer and normal groups. In contrast, the patterns generated by the cancer and normal groups looked highly similar to each other.

Furthermore, the patterns detected in the first benign-tumour group looked almost identical to the patterns detected in all three groups in the second data set — cancers, normals and patients with benign tumours. This indicated to Baggerly that Liotta and his colleagues inadvertently changed their experimental set-up midway through collecting the first data set — the one that they analysed for their *Lancet* paper.

Although the second and third data sets seemed consistent with each other, the precise differences in protein patterns that separated cancers from controls in one set could not be used to make the same distinction in the other — which makes little sense if the test is uncovering fundamental biological differences between cancerous and control samples. Baggerly's team also claimed that the data were collected on a machine that had not been properly calibrated, or adjusted for accuracy. In a paper first published online in January this year[3], the researchers concluded that the test may be uncovering differences due to "artifacts of sample processing, not to the underlying biology of cancer".

Divided over data

Baggerly says that he discussed his concerns privately with Liotta and Petricoin in December 2002. He decided to publish only after learning that a test might become commercially available.

Petricoin says there is no proof that m/z values below 2,000 always represent experimental noise or bias. He also denies that the team switched methods midway through the first experiment. And he explains that the second and third data sets were processed differently. "The point is, the *Lancet* paper shows feasibility of this approach, and the results derived from each of these data sets prove that there do indeed exist low-molecular-weight molecules in the circulation that can discriminate the disease states," says Petricoin, who adds that the team has further refined its methods in a new paper[4]. He also notes that other groups have examined his data and supported his conclusions[5].

Petricoin stresses that he and Liotta are not directly involved with commercial development of the OvaCheck test. Peter Levine, Correlogic's chief executive, says that the company has also refined its data analysis techniques since the *Lancet* paper. "It seems to me a lot of people are sort of debating an issue that is pretty much of historical relevance only," he says. But Correlogic has not released its data, so it is impossible to verify this claim.

Meanwhile, Correlogic is now in dispute with Liotta and Petricoin over the pair's consulting work for a rival company, Biospect of South San Francisco. On 18 May, the two scientists were called before a congressional committee and asked whether this had slowed OvaCheck's development — a charge that they roundly denied.

Growing pains

But this dispute is secondary to the main issue of whether the technology works. Sceptics want to see more evidence that it yields consistent results on samples from different labs. And some argue that the field needs to develop standards for how proteomics experiments should be done and reported.

Because the technology is so new, says epidemiologist David Ransohoff of the University of North Carolina in Chapel Hill, scientists are still learning how to cut out all the possible sources of bias in proteomics experiments. For instance, if the cancer samples are collected from women being tested because they are known to be at high risk of suffering from the disease, they might experience anxiety when sampled, unlike controls whose samples may be taken as part of a routine check-up. In this case, the first group of samples may be flooded with stress hormones that would be detected by a proteomic analysis, but have nothing to do with whether or not a woman has cancer.

"Whether the test works or not, we will learn from this experience with OvaCheck what rules of evidence we might apply in the future to find useful results more efficiently," says Ransohoff.

He hopes that the episode will lead to an established set of standards for evaluating the effectiveness of such tests. Only then will we know whether proteomics-based diagnostic tools truly deserve the trust of scientists, doctors — and, most importantly, patients. ■

Erika Check is *Nature*'s Washington biomedical correspondent.

1. Petricoin, E. F. *et al. Lancet* **359,** 572–577 (2002).
2. Sorace J. M. & Zhan, M. *BMC Bioinformatics* **4,** 24 (2003).
3. Baggerly, K. A., Morris, J. S. & Coombes, K. R. *Bioinformatics* **20,** 777–785 (2004).
4. Conrads, T. P. *et al. Endocr.-Relat. Cancer* **11,** 163–178 (2004).
5. Zhu, W. *et al. Proc. Natl Acad. Sci. USA* **100,** 14666–14671 (2003).

http://ncifdaproteomics.com