measurements. In absorption spectroscopy, the signal is sensed only indirectly, from the light that does not interact with the sample (Fig. 1a). Weak absorption is therefore very difficult to detect, because it changes the intensity of the transmitted light only marginally. Theoretically, the detection of weak absorbers could be improved by increasing the intensity of the incident light, but commonly used infrared detectors become less sensitive at higher light intensities[12], imposing a practical limit on the maximum light intensity that can be used. By contrast, Pupeza *et al.* detect the signal of interest — the radiation emitted from the vibrating molecules — directly (Fig. 1b). This is analogous to the difference between absorbance and fluorescence measurements in the visible spectral range: fluorescence measurements are the more sensitive because they detect a signal directly from the sample, and can even detect it from a single molecule.

Pupeza and colleagues demonstrate the high sensitivity of their approach in various ways. For example, they were able to detect 40-fold lower concentrations of a compound in solution, and to better distinguish between two similar compounds, than when using absorption spectroscopy. They also obtained spectra of biological samples that block nearly all of the incoming light (in one case, at least 99.999%). Thus, the new approach senses light where currently used methods see only darkness. This is an impressive achievement, and might alleviate both of the main problems of conventional infrared spectroscopy: sensitivity and strong infrared absorption by water. It will simplify sample preparation in many cases by removing the need for sample concentration or drying, and will open up new applications — particularly those involving aqueous biological samples.

The authors suggest several ideas for taking the method further, such as by increasing the power of the laser used to irradiate the sample. It is to be hoped that such measures will further narrow the technological gap that at present prevents the method from achieving the ultimate goal of single-molecule sensitivity in bulk water. Other challenges will be to increase the spectral range of the measurements to include the shorter wavelengths at which prominent and diagnostically useful signals are found for proteins, lipids and nucleotides, and to develop a spectrometer suitable for commercialization at a competitive price.

Andreas Barth is in the Department of Biochemistry and Biophysics, Stockholm University, Stockholm 106 91, Sweden.
e-mail: barth@dbb.su.se

1. Pupeza, I. *et al. Nature* **577**, 52–59 (2020).
2. Herschel, W. *Phil. Trans. R. Soc. Lond.* **90**, 284–292 (1800).
3. van Dishoeck, E. F. *Annu. Rev. Astron. Astrophys.* **42**, 119–167 (2004).
4. Ehrenfreund, P. & Charnley, S. B. *Annu. Rev. Astron. Astrophys.* **38**, 427–483 (2000).
5. Bibring, J.-P. *et al. Nature* **428**, 627–630 (2004).
6. Chalmers, J. M. & Griffiths, P. R. (eds) *Handbook of Vibrational Spectroscopy* Vols 4 & 5 (Wiley, 2001).
7. Barth, A. & Haris, P. I. (eds) *Biological and Biomedical Infrared Spectroscopy* (IOS, 2009).
8. Sun, D.-W. (ed.) *Infrared Spectroscopy for Food Quality Analysis and Control* (Academic, 2008).
9. Chalmers, J. M., Edwards, H. G. M. & Hargreaves, M. D. (eds) *Infrared and Raman Spectroscopy in Forensic Science* (Wiley, 2012).
10. Bunaciu, A. A., Fleschin, Ş., Hoang, V. D. & Aboul-Enein, H. Y. *Crit. Rev. Anal. Chem.* **47**, 67–75 (2017).
11. Wu, Q. & Zhang, X.-C. *Appl. Phys. Lett.* **67**, 3523–3525 (1995).
12. Theocharous, E., Ishii, J. & Fox, P. N. *Appl. Opt.* **43**, 4182–4188 (2004).

Medical research

AI shows promise for breast cancer screening

Etta D. Pisano

Could artificial intelligence improve the accuracy of screening for breast cancer? A comparison of the diagnostic performance of expert physicians and computers suggests so, but the clinical implications are as yet uncertain. **See p.89**

Screening is used to detect breast cancer early in women who have no obvious signs of the disease. This image-analysis task is challenging because cancer is often hidden or masked in mammograms by overlapping 'dense' breast tissue. The problem has stimulated efforts to develop computer-based artificial-intelligence (AI) systems to improve diagnostic performance. On page 89, McKinney *et al.*[1] report the development of an AI system that outperforms expert radiologists in accurately interpreting mammograms from screening programmes. The work is part of a wave of studies investigating the use of AI in a range of medical-imaging contexts[2].

Despite some limitations, McKinney and colleagues' study is impressive. Its strengths include the large scale of the data sets used for training and subsequently validating the AI algorithm. Mammograms for 25,856 women in the United Kingdom and 3,097 women in the United States were used to train the AI system. The system was then used to identify the presence of breast cancer in mammograms of women who were known to have had either biopsy-proven breast cancer or normal follow-up imaging results at least 365 days later. These outcomes are the widely accepted gold standard for confirming breast cancer status in people undergoing screening for the disease. The authors report that the AI system outperformed both the historical decisions made by the radiologists who initially assessed the mammograms, and the decisions of 6 expert radiologists who interpreted 500 randomly selected cases in a controlled study.

McKinney and colleagues' results suggest that AI might some day have a role in aiding the early detection of breast cancer, but the authors rightly note that clinical trials will be needed to further assess the utility of this tool in medical practice. The real world is more complicated and potentially more diverse than the type of controlled research environment reported in this study. For example, the study did not include all the different mammography technologies currently in use, and most images were obtained using a mammography system from a single manufacturer. The study included examples of two types of mammogram: tomosynthesis (also known as 3D mammography) and conventional digital (2D) mammography. It would be useful to know how the system performed individually for each technology.

> **"Clinical trials will be needed to further assess the utility of this tool in medical practice."**

The demographics of the population studied by the authors is not well defined, apart from by age. The performance of AI algorithms can be highly dependent on the population used in the training sets. It is therefore important that a representative sample of the general population be used in the development of this technology, to ensure that the results are broadly applicable.

Another reason to temper excitement about this and similar AI studies is the lessons learnt from computer-aided detection (CAD) of breast cancer. CAD, an earlier computer system aimed at improving mammography interpretation in the clinic, showed great promise in experimental testing, but fell short in real-world settings[3]. CAD marks

mammograms to draw the interpreter's attention to areas that might be abnormal. However, analysis of a large sample of clinical mammography interpretations from the US Breast Cancer Surveillance Consortium registry demonstrated that there was no improvement in diagnostic accuracy with CAD[3]. Moreover, that study revealed that the addition of CAD worsened sensitivity (the performance of radiologists in determining that cancer was present), thus increasing the likelihood of a false negative test. CAD did not result in a significant change in specificity (the performance of radiologists in determining that cancer was not present) and the likelihood of a false positive test[3].

It has been speculated that CAD was not as useful in the clinic as experimental data suggested it might be because radiologists ignored or misused its input owing to the high frequency of marks on the images that were not findings suggestive of cancer. This outcome was attributed by some to the limited processing power available for CAD, which meant that comparisons with previous imaging studies of the same person were not possible[4]. Thus, CAD might mark regions that were not changing over time and that could be easily dismissed by expert readers. Another factor that limited CAD is that it was developed using the performance of human-based diagnosis. It was trained using mammograms in which humans had found signs of cancer and others that were false negatives – cases in which humans could not see signs of cancer although the disease was indeed present[4]. Similar pitfalls could be encountered with AI-based decision aids, too.

A system by which AI finds abnormalities that humans miss will require radiologists to adapt to the use of these types of tool. Imagine a system in which an algorithm marks a dense breast area on a screening mammogram and the human radiologist cannot see anything that looks potentially malignant. With CAD, radiologists scrutinize the areas marked, and if they decide the mark is probably not cancer, they assign the mammogram as being negative for malignancy. However, if AI algorithms are to make a bigger difference than CAD in detecting cancers that are currently missed, an abnormality detected by the AI system, but not perceived as such by the radiologist, would probably require extra investigation. This might result in a rise in the number of people who receive callbacks for further evaluation. A clinical trial would show the effect of the AI system on the detection of cancer and the rate of false positive diagnoses, while also allowing the development of effective clinical practice in response to mammograms flagged as abnormal by AI but not by the radiologist.

In addition, it would be essential to develop a mechanism for monitoring the performance of the AI system as it learns from cases it encounters, as occurs in machine-learning algorithms. Such performance metrics would need to be available to those using these tools, in case performance deteriorates over time.

It is sobering to consider the sheer volume of data needed to develop and test AI algorithms for clinical tasks. Breast cancer screening is perhaps an ideal application for AI in medical imaging because large curated data sets suitable for algorithm training and testing are already available, and information for validating straightforward clinical end points is readily obtainable. Breast cancer screening programmes routinely measure their diagnostic performance – whether cancer is correctly detected (a true positive) or missed (a false negative). Some areas found on mammograms might be identified as abnormal but turn out on further testing not to be cancerous (false positives). For most women, screening identifies no abnormalities, and when there is still no evidence of cancer one year later, this is classified as a true negative.

Most other medical tasks have more-complicated clinical outcomes, however, in which the clinician's decision is not a binary one (between the presence or absence of cancer), and thus further signs and symptoms must also be considered. In addition, most diseases lack readily accessible, validated data sets in which the 'truth' is defined relatively easily. Obtaining validated data sets for more-complex clinical problems will require greater effort by readers and the development of tools that can interrogate electronic health records to identify and annotate cases representing specific diagnoses.

To achieve the promise of AI in health care that is implied by McKinney and colleagues' study, anonymized data in health records might thus have to be treated as precious resources of potential benefit to human health, in much the same way as public utilities such as drinking water are currently treated. Clearly, however, if such AI systems are to be developed and used widely, attention must be paid to patient privacy, and to how data are stored and used, by whom, and with what type of oversight.

Etta D. Pisano is at the American College of Radiology, Philadelphia, Pennsylvania 19103, USA, and at Beth Israel Lahey Medical Center, Harvard Medical School, Boston, Massachusetts.
e-mail: episano@bidmc.harvard.edu

1. McKinney, S. M. *et al. Nature* **577**, 89–94 (2020).
2. Neri, E. *et al. Insights Imaging* **10**, 44 (2019).
3. Lehman, C. D. *et al. JAMA Intern. Med.* **175**, 1828–1837 (2015).
4. Kohli, A. & Jha, S. *J. Am. Coll. Radiol.* **15**, 535–537 (2018).

Astronomy

Galaxy cluster illuminates the cosmic dark ages

Nina A. Hatch

Observations of a distant cluster of galaxies suggest that star formation began there only 370 million years after the Big Bang. The results provide key details about where and when the first stars and galaxies emerged in the Universe. **See p.39**

Shortly after the Big Bang, the Universe was completely dark. Stars and galaxies, which provide the Universe with light, had not yet formed, and the Universe consisted of a primordial soup of neutral hydrogen and helium atoms and invisible 'dark matter'. During these cosmic dark ages, which lasted for several hundred million years, the first stars and galaxies emerged. Unfortunately, observations of this era are challenging because dark-age galaxies are exceptionally faint[1]. On page 39, Willis *et al.*[2] provide a glimpse of what happened during the dark ages by doing some galactic archaeology. By measuring the ages of stars in one of the most distant clusters of galaxies known, the authors located galaxies that formed stars in the dark ages, close to the earliest possible time that stars could emerge.

A galaxy cluster is a group of thousands of galaxies that orbit each other at speeds[3] of about 1,000 kilometres per second. They are prevented from flying apart by the gravitational pull of the accompanying dark matter, which has the equivalent total mass of about one hundred trillion Suns[4]. Astronomers use these clusters as laboratories for many experiments in astrophysics, such as measuring the composition of the Universe, testing theories of gravity and determining how galaxies form. Willis *et al.* used one of the