research update

FROM THE JOURNALS Edited highlights of Richard Lehman's blog on http://bmj.co/Lehman

Screening for bowel cancer

This is a good week to sharpen your understanding of the issues around bowel cancer screening. Two editorials from people outside the gastroenterology/screening community take a critical look at what's going on in the United States. Rita Redberg's piece (*JAMA Intern Med* 2016, doi:10.1001/jamainternmed.2016.3892) could not be clearer.

Despite a gut feeling (pardon the expression) that colonoscopy has to be better than flexible sigmoidoscopy, there's no evidence for that, and we do know that it can cause more complications.

Faecal occult blood testing is pretty hit and miss as an initial population screening stool—sorry, I mean tool; but so is a new blood test that is being touted in the United States. This is the subject of an editorial by Ravi Parikh and Vinay Prasad on the *JAMA* website (doi:10.1001/jama.2016.7914). Both articles are well worth going through. The evidence that bowel cancer screening saves lives (even on a disease specific level) is remarkably flimsy.

SPL

Let them take letrozole

"The MA.17R trial was a phase 3, randomized, double-blind, placebo-controlled trial involving postmenopausal women with primary breast cancer who had received 4.5 to 6 years of adjuvant therapy with an aromatase inhibitor, preceded in most patients by treatment with tamoxifen. Within 2 years after completing treatment with the aromatase inhibitor, patients were randomly assigned to receive 2.5 mg of letrozole or placebo orally once a day for another 5 years." The benefits of letrozole consisted of a 4% absolute difference in recurrence-free survival and a 0.28% difference in contralateral breast cancer, both on the edge of statistical significance. Overall survival did not differ. On the harms side, there was a 6% greater chance of bone fractures and osteoporosis when taking letrozole. How much more useful this paper would have been if it had included a decision aid to be shared between doctors and patients.

N Engl J Med 2016, doi:10.1056/NEJMe1606031

Lower systolic blood pressure no better in brain bleeds

Intensive Blood-Pressure Lowering In Patients With Acute Cerebral Haemorrhage. Thank you *New England Journal of Medicine* for publishing a trial of real clinical importance, called ATACH-2. It was stopped early for futility. Clinicians can content themselves with a systolic blood pressure target between 140 and 179 mm Hg: a lower target brings no gain in lives or function.

N Engl J Med 2016, doi:10.1056/NEJMoa1603460

Aspirin for ARDS: read and forget

When I read the title Effect of Aspirin on Development of ARDS in At-Risk Patients Presenting to the Emergency Department I thought, help, I'll have to get my head round this. Acute respiratory distress syndrome is something I've never witnessed under that label, which I think it acquired about 20 years ago. I imagined that it was a multifactorial and unpredictable physiological syndrome, but here the researchers seemed able to identify at risk patients. What were the selection criteria and could aspirin really have an effect on the syndrome? I need not have worried. Aspirin had no effect, so I can leave the matter alone and move on.

JAMA 2016, doi:10.1001/jama.2016.6330

Opioids for chronic pain can kill

The popularity of long acting opioids for non-cancer pain in North America is undoubtedly leading to thousands of deaths. Quantifying them, however, is a tricky business. Here's an observational study from Tennessee that tries to use propensity scoring to allow comparisons between new episodes of prescribed treatment for long acting opioids and either analgesic anticonvulsants or low dose cyclic antidepressants (control drugs). Allowing for time differences between groups, those prescribed opioids had at least a 64% higher mortality than those prescribed non-opioids in the first few months. In fact, it was more than 90% in the case of out-of-hospital deaths, and this was only partly explained by unintentional overdosage. Here's a situation where you could quibble about confounding and absolute risk differences and speculate about causation, but the clear fact is that these drugs are dangerous.

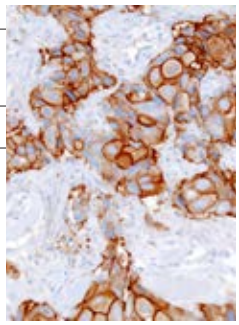
JAMA 2016, doi:10.1001/jama.2016.7789

Safe drugs to help quit smoking

It's always nice when a randomised trial confirms what observational evidence has already found. Population-wide linkage studies have shown no evidence of neuropsychiatric harm from using bupropion or varenicline as an aid to smoking cessation. The triple dummy EAGLES trial confirms this. Also, "Varenicline was more effective than placebo, nicotine patch, and bupropion in helping smokers achieve abstinence, whereas bupropion and nicotine patch were more effective than placebo." All good to know, but it seems to me that a government serious about protecting its citizens from the harms of combustible tobacco would simply make the stuff unavailable. E-cigarettes and safer forms of nicotine are everywhere. But don't hold your breath: doing this would cost £9bn in tax revenue.

Lancet 2016, doi:10.1016/S0140-6736(16)30272-0

Role of second opinions for breast histopathology



ORIGINAL RESEARCH Simulation study

Evaluation of 12 strategies for obtaining second opinions to improve interpretation of breast histopathology

Elmore JG, Tosteson ANA, Pepe MS, et al

Cite this as: *BMJ* 2016;353:i3069
Find this at: http://dx.doi.org/10.1136/bmj.i3069

Study question What effect do second opinions have on improving accuracy of interpreting breast histopathology?

Methods Interpretations from 115 pathologists, one slide for each case, were used to establish baseline accuracy of single observers. These interpretations were compared with accuracy based on independent interpretations by pairs of pathologists, with resolution by an independent third pathologist when needed. The researchers evaluated 12 strategies, with acquisition of second opinions dependent on initial diagnoses, assessment of case difficulty or borderline diagnostic characteristics, pathologists' clinical volumes of breast biopsy specimens, or whether a second opinion was required by policy or desired by the pathologists. The researchers compared the diagnoses (initial and after the second opinion) with expert consensus-derived reference diagnoses to calculate rates of misclassification; agreement statistics were calculated for different assessments of the same case. The 240 cases included benign without atypia (10% non-proliferative, 20% proliferative without atypia), atypia (30%), ductal carcinoma in situ (DCIS, 30%), and invasive cancer (10%).

Study answer and limitations Misclassification rates significantly decreased (P<0.001) with all second opinion strategies except for the strategy limiting second opinions only to initial diagnoses of invasive cancer. The overall misclassification rate decreased from 24.7% to 18.1% when all cases received second opinions (P<0.001). The lowest misclassification rate in this test set resulted when high volume pathologists provided both first and second opinions (14.3%, 95% confidence interval 10.9% to 18.0%). Obtaining second opinions only for cases with initial interpretations of atypia, DCIS, or invasive cancer decreased the over-interpretation of benign cases without atypia from 12.9% to 6.0%. These statistics depend on the composition of the test set, which included a higher prevalence of difficult cases than are seen in typical practice. Atypia cases had the highest misclassification rate after single interpretation (52.2%), remaining at more than 34% for all second opinion scenarios.

What this study adds Systematic application of second opinions can significantly improve the accuracy of interpreting breast histopathology but will not completely eliminate diagnostic variability, especially for breast biopsy specimens with atypia.

Funding, competing interests, data sharing Funded by the National Cancer Institute. Authors should be contacted for data sharing.

COMMENTARY Diagnostic accuracy improves when pathologists work together

Most pathologists do not work alone. Because of sheer volume and complexity of interpretation, breast biopsy specimens often receive secondary review. This presumably reflects the belief that another opinion can minimise or eliminate uncertainty and ultimately improve patient outcome.

In this issue, Elmore and colleagues report a simulation study to compare 12 different strategies for obtaining second opinions to improve the interpretation of breast histopathology.[1] Strategies ranged from second opinion for all case to selective strategies based on initial diagnosis, pathologist desire and perception of difficulty, and institutional policy.

Take two

Misclassification rates decreased by 6.6% when all types of case had a second opinion, whereas second opinions for all cases with initial interpretation other than benign led to a 6.9% reclassification rate. The lowest rate of misclassification was seen when pathologists who dealt with a high volume of breast specimens provided both first and second opinions.

> It seems highly likely that second opinions in breast pathology safeguard patients against over-treatment or under-treatment

Although this study is a simulation, the results and conclusions are consistent with other published studies.[6 7] Changes in interpretation usually occur when smaller laboratories in the community, staffed by general pathologists, send specimens to larger academic laboratories, staffed by specialty pathologists. A limitation of all these studies is a lack of evidence that these changes in interpretation improve patient outcomes; such a study would be impossible as the clinical course of many breast lesions is altered by their excision. Nevertheless, it seems highly likely that second opinions in breast pathology safeguard patients against over-treatment or under-treatment.

However, the nature of a useful second opinion is likely to vary among different settings. A pathologist working in a small practice or small community hospital is likely to benefit from sending borderline, difficult, or unusual cases to a larger volume more specialised academic centre. The large academic centres benefit from ongoing quality assurance programmes, such as random review of 5-10% of cases, double reads of all breast core biopsy specimens, or at least atypical and malignant diagnoses, slide review at multidisciplinary treatment conferences, and review of interesting, difficult, and challenging cases at daily pathology departmental conferences.

Elmore and colleagues' study provides quantitative data that will refine the already common practice of seeking second opinions in pathology.[9] Any quality assurance protocol designed to reach concordance and reduce diagnostic errors will ultimately benefit the patients that we all serve. As this study shows, there is still plenty of room for improvement.

Nancy E Davidson davidsonne@upmc.edu
See thebmj.com for author details

Cite this as: *BMJ* 2016;353:i3256
Find this at: http://dx.doi.org/10.1136/bmj.i3256

ORIGINAL RESEARCH Propensity weighted nationwide cohort study

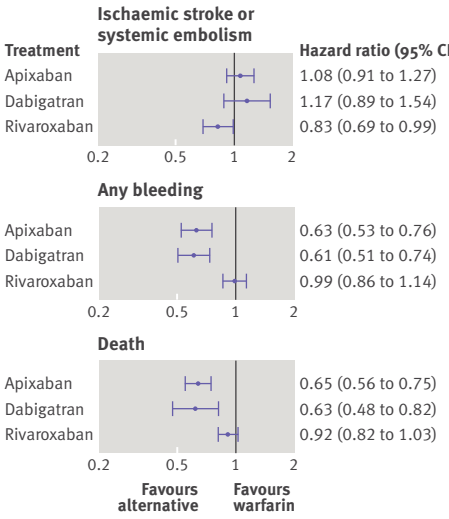
Comparative effectiveness and safety of NOACs and warfarin in patients with atrial fibrillation

Larsen TB, Skjøth F, Nielsen PB, Kjældgaard JN, Lip GYH

Cite this as: *BMJ* 2016;353:i3189
Find this at: http://dx.doi.org/10.1136/bmj.i3189

Study question How effective and safe are the non-vitamin K antagonist oral anticoagulants (NOACs) dabigatran, rivaroxaban, and apixaban compared with warfarin?

Methods The cohort in this Danish study comprised anticoagulant naive patients who had started treatment for atrial fibrillation from August 2011 to October 2015. Eligible participants had no previous indication for valvular atrial fibrillation nor venous thromboembolism. The authors evaluated effectiveness (stroke) and safety outcomes (bleeding) in patients who started treatment with a NOAC compared with those who started treatment with warfarin.



Propensity weighted Cox hazard ratios for NOACs compared with warfarin for effectiveness and safety endpoints

Study answers and limitations The study population (n=61 678) was distributed according to treatment as: warfarin (2.5 mg variable dose) n=35 436 (57%), dabigatran (150 mg twice daily) n=12 701 (21%), rivaroxaban (20 mg once daily) n=7192 (12%), and apixaban (5 mg twice daily) n=6349 (10%). During one year's follow-up, rivaroxaban was associated with lower rates of ischaemic stroke or systemic embolism (3.0%) compared with warfarin (3.3%) with a hazard ratio of 0.83 (95% confidence interval 0.69 to 0.99). The hazard ratios for dabigatran (2.8%) and apixaban (4.9%) were non-significant compared with those for warfarin. When the analysis was restricted to ischaemic stroke, NOACs were not significantly different from warfarin. Annual bleeding rates for apixaban (3.3%) and dabigatran (2.4%) were significantly lower than for warfarin (5.0%) (0.62, 0.51 to 0.74). The rates for warfarin and rivaroxaban (5.3%) were comparable. The results may be biased because of unobserved residual confounding and selective prescribing behaviour.

What this study adds Our findings provide reassurance that NOACs are safe and effective alternatives to warfarin in a routine care setting.

Funding, competing interests, data sharing Funded partly by an unrestricted grant from the Obel Family Foundation. Competing interests are available in the full paper on bmj.com. Data sharing is not possible.

ORIGINAL RESEARCH Comparative observational study of benefits and harms

Prolonged dual antiplatelet therapy in stable coronary disease

Timmis A, Rapsomaniki E, Chung S C, et al

Cite this as: *BMJ* 2016;353:i3163
Find this at: http://dx.doi.org/10.1136/bmj.i3163

Study question What is the potential magnitude of benefits and harms in trial findings for extended dual antiplatelet therapy with aspirin and ticagrelor in unselected patients who survive a year or more after acute myocardial infarction?

Methods A trial population (PEGASUS-TIMI-54) was compared with an observational population based cohort of 7238 patients by using linked primary and secondary care electronic health records (CALIBER (ClinicAl research using LInked Bespoke studies and Electronic health Records)).

Study answer and limitations 23.1% of the CALIBER cohort comprised the "target" population that met trial inclusion and exclusion criteria. Compared with the placebo arm in the trial population, in the target population the median age was 12 years higher, there were more women (48.6% *v* 24.3%), and there was a substantially higher cumulative three year risk of both the primary (benefit) trial endpoint of recurrent acute myocardial infarction, stroke, or fatal cardiovascular disease (18.8% (95% confidence interval 16.3% to 21.8%) *v* 9.04%) and the primary (harm) endpoint of fatal, severe, or intracranial bleeding (3.0% (2.0% to 4.4%) *v* 1.26%)).

Observed cumulative event rate, and number of events prevented or harms caused (with 95% CI) applying PEGASUS-TIMI-54 trial results to CALIBER target population that met trial inclusion and exclusion criteria

	CALIBER target population (n=1676)	PEGASUS-TIMI-54 trial placebo arm (n=7067)
MI/stroke/fatal cardiovascular disease		
3 year cumulative risk (%)	18.8 (16.3 to 21.8)	9.04
Events prevented/year/10 000 patients treated*	101 (87 to 117)	—
Fatal, severe, or intracranial bleeding		
3 year cumulative risk (%)	3.0 (2.0 to 4.4)	1.26
Excess harms/year/10 000 patients treated*	75 (50 to 110)	—

*Calculated by applying risk reduction in PEGASUS-TIMI-54 trial.

Application of intention to treat relative risks from the trial (ticagrelor 60 mg daily arm) to CALIBER's target population showed an estimated 101 (95% confidence interval 87 to 117) ischaemic events prevented per 10 000 patients treated per year and an estimated 75 (50 to 110) excess fatal, severe, or intracranial bleeds caused per 10 000 patients treated per year. In the CALIBER cohort, a precise match was available for only two of the trial's major bleeding criteria.

What this study adds This novel use of primary-secondary care linked electronic health records allows characterisation of "healthy trial participant" effects and provides estimates of the potential absolute benefits and harms of dual antiplatelet therapy in representative patients who survived a year or more after acute myocardial infarction. The methods are scalable and could provide useful real world evidence for trial results in other clinical settings.

Funding, competing interests, data sharing This study was supported by AstraZeneca, the National Institute for Health Research, Wellcome Trust, and the Medical Research Council, and by awards from multiple funders to establish the Farr Institute of Health Informatics Research at UCL. CE is employed by AstraZeneca. There are no data to share.

A fresh look at risk prediction

RESEARCH METHODS AND REPORTING Opportunities and challenges of "big" datasets

External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis

Riley RD, Ensor J, Snell KIE, et al

Cite this as: *BMJ* 2016;353:i3140
Find this at: http://dx.doi.org/10.1136/bmj.i3140

Clinical prediction models are used to predict the risk of disease presence and outcome occurrence in individuals, thereby informing clinical diagnosis and prognosis. Unfortunately, most prediction research focuses on model development and there are relatively few external validation studies. External validation uses new participant level data, external to that used for model development, to examine whether the model's predictions are reliable in individuals from potential population(s) for clinical use. It is crucial, because developed models are often produced in selected populations and over-fitted to the observed data, leading to over-optimism in their apparent performance.

A shortage of external validation studies is often attributed to the lack of data available besides those used for model development. However, increasingly researchers have access to so-called "big" datasets, as shown by meta-analyses using individual participant data (IPD) from multiple studies, and by analyses of registry databases containing electronic health (e-health) records for thousands or even millions of patients. For example, QRISK2, which estimates cardiovascular risk, was developed using e-health data with over 1.5 million patients (with over 95 000 new cardiovascular events) from 355 randomly selected general practices, and externally validated in an additional 1.6 million patients from a further 365 practices.

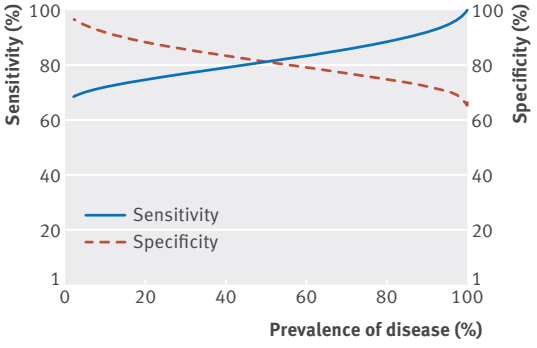
Such big datasets herald an exciting opportunity to improve the uptake and scope of external validation research. This article describes the additional opportunities, challenges, and reporting issues in this situation. In particular, researchers are encouraged to use big datasets to externally validate a model's predictive performance across all clinical settings, populations, and subgroups of interest. Simply reporting a model's overall performance (averaged across all individuals and populations) is not sufficient, because it can mask differences and important deficiencies. The article's examples show how a model often needs tailoring or updating to perform well in particular subgroups and settings, and reinforce why data sharing should be the expected norm, to facilitate such detailed interrogation.

RESEARCH METHODS AND REPORTING Analysis of the spectrum effect

The spectrum effect in tests for risk prediction, screening, and diagnosis

Usher-Smith JA, Sharp SJ, Griffin SJ

Cite this as: *BMJ* 2016;353:i3139
Find this at: http://dx.doi.org/10.1136/bmj.i3139



Variation in sensitivity (blue solid line) and specificity (red dashed line) with true prevalence of a disease, where true prevalence is changed by varying the mean of a normal distribution of a continuous variable X while keeping the threshold value used to define disease constant

Much of clinical practice relies on tests measuring one or more characteristics of an individual to determine whether that individual is at risk of developing or does or does not have a particular disease. The spectrum effect describes the variation in test performance among different population subgroups. It arises because most situations in clinical practice have a continuum of characteristics on which the classification of disease status is based, and all tests are subject to error. The result is that the performance of tests is influenced by both the prevalence of the disease in the sample and the characteristics of the sample.

These effects can be illustrated by simulating a situation in which there is a continuous variable X with a value for each individual, and true disease status is defined by the value of X being above or below a particular threshold (for example, fasting glucose for which the disease is diabetes and the threshold for X is 7.0 mmol/L). When prevalence is varied by changing the mean of the normal distribution of X while keeping the threshold constant for example, as prevalence decreases, the sensitivity and specificity vary by 30% with sensitivity decreasing and specificity increasing (figure).

Tests developed in populations with a higher prevalence of disease will, therefore, typically have a lower sensitivity and higher specificity when applied in populations with lower disease prevalence. Calculations of positive and negative predictive value will only in part adjust for differences in disease prevalence. When reviewing a study of a new risk prediction, screening, or diagnostic test and deciding whether to use that test in practice, clinicians and policymakers should, therefore, examine the relevance of the study sample to their own population. Ideally new tests should be developed and evaluated using data from the population(s) in which they are intended to be used.