BMJ

BMJ 2013;346:e5793 doi: 10.1136/bmj.e5793 (Published 5 February 2013)

RESEARCH METHODS & REPORTING

Prognosis research strategy (PROGRESS) 4: Stratified medicine research

In patients with a particular disease or health condition, stratified medicine seeks to identify those who will have the most clinical benefit or least harm from a specific treatment. In this article, the fourth in the PROGRESS series, the authors discuss why prognosis research should form a cornerstone of stratified medicine, especially in regard to the identification of factors that predict individual treatment response

Aroon D Hingorani *professor of genetic epidemiology*[1], Daniëlle A van der Windt *professor in primary care epidemiology*[2], Richard D Riley *senior lecturer in medical statistics*[3], Keith Abrams *professor of medical statistics*[4], Karel G M Moons *professor of clinical epidemiology*[5], Ewout W Steyerberg *professor of medical decision making*[6], Sara Schroter *senior researcher*[7], Willi Sauerbrei *professor of medical biometry*[8], Douglas G Altman *professor of statistics in medicine*[9], Harry Hemingway *professor of clinical epidemiology*[1], for the PROGRESS Group

[1]Department of Epidemiology and Public Health, University College London, London WC1E 7HB, UK; [2]Arthritis Research UK Primary Care Centre, Primary Care Sciences, Keele University, Keele ST5 5BG, UK; [3]School of Health and Population Sciences, University of Birmingham, Birmingham B15 2TT, UK; [4]Centre for Biostatistics & Genetic Epidemiology, Department of Health Sciences, School of Medicine, University of Leicester, Leicester LE1 7RH, UK; [5]Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht, Netherlands; [6]Department of Public Health, Erasmus MC, 3000 CA Rotterdam, Rotterdam, Netherlands; [7]BMJ, BMA House, London WC1H 9JR, UK; [8]Institute of Medical Biometry and Informatics, University Medical Center Freiburg, 79104 Freiburg, Germany; [9]Centre for Statistics in Medicine, University of Oxford, Oxford OX2 6UD, UK

A woman with newly diagnosed breast cancer is deciding on a course of therapy, guided by her physician. Evidence on the average prognosis[1] and effectiveness of therapeutic interventions is available from studies of large groups of patients with breast cancer in observational studies and randomised trials. But the patient and doctor are faced with making a decision in an individual case, where the prognosis and response to treatment may deviate from average. One way to select the optimal treatment is to consider a test that predicts treatment effect, such as the human epidermal growth factor receptor 2 (HER-2) status.[2] The use of HER-2 status in breast cancer management is an example of the translation of results from prognosis research toward improved patient outcomes. The prognosis of breast cancer patients is highly variable,[1] HER-2 was discovered as a prognostic factor,[3] which provided a specific target for an intervention (trastuzumab), which was then evaluated in trials which recruited women with HER-2 positive cancers (see fig 1⇓). After the success of these trials in improving clinical outcome, trastuzumab is now given to the subgroup (stratum) of women who are HER-2 positive, but not to those testing negative;[4] this type of approach has been termed stratified medicine.

The aims of this fourth paper in our PROGRESS series (www.progress-partnership.org) are to describe the rationale for stratified medicine, and to explain why prognosis research is pivotal for this purpose; from identifying priority areas for stratification, to discovering candidate factors that may predict treatment response, through to trials and health technology assessment that examine the impact of stratified medicine approaches in healthcare. We identify current challenges and deficiencies in such research and make recommendations for improvement with examples across a variety of disease areas.

What is stratified medicine?

Stratified medicine refers to the targetting of treatments (including pharmacological and non-pharmacological

Correspondence to: H Hemingway h.hemingway@ucl.ac.uk

Extra material, as supplied by the author (see http://www.bmj.com/content/346/bmj.e5793?tab=related#webextra)

Supplementary table: Recommendations of PROGRESS (PROGnosis RESearch Strategy)

interventions) according to the biological or risk characteristics shared by subgroups of patients. Stratified medicine is regarded as central to the progress of healthcare according to the leaders of the National Institutes of Health, and the Food and Drug Administration[6] among others.[7] In contrast with "all comer" or "empirical" medicine, stratified medicine seeks to target therapy and make the best decisions for groups of similar patients.[8 9]

One approach to stratifying the use of treatments is to consider absolute risks. In the third article of our series[10] we described how prognostic models are used to estimate the absolute risk of an outcome for an individual. Those people with the highest absolute risk will derive the largest absolute benefit from a treatment (that is, the greatest reduction in probability of the outcome) when the treatment effect expressed in relative terms is the same for all patients. This is illustrated in the upper panel of fig 2⇓, where the relative treatment effect on mortality risk is estimated as 0.75 for all patients but the reduction in absolute probability of death is 5% for low risk patients and 15% for high risk patients. In such situations treatments could be restricted (or "personalised") to those who will benefit the most. Examples in common clinical practice include the decision to give lipid lowering therapy to people above a certain threshold of cardiovascular risk estimated from a prognostic model,[11] the use of bisphosphonates for women over the age of 50 considered to have an increased risk of vertebral fractures, and the targeting of primary care management of back pain.[12]

By contrast, clinicians may also stratify medicine because the relative treatment effect is inconsistent across patients (fig 2, lower panel⇓). In this situation, at least one individual patient measure is associated with changes in the treatment effect. In statistical terms there is an interaction between a patient-level variable and the effect of treatment on the outcome, and in biological terms there may be an underlying mechanism explaining the interaction. In this situation, a stratified medicine approach seeks to test patients for the presence of individual factors that are considered predictive of an improved treatment response (more benefit, less harm, or both), as in the aforementioned test for positive HER-2 status in breast cancer and the use of trastuzumab. Other examples in clinical use include imatinib in patients with chronic myeloid leukaemia targeted to those with the BCR-ABL mutation[13] and gefitinib used to treat pulmonary adenocarcinoma in patients with epidermal growth factor receptor mutations.[14]

An example of identifying patients with greater risk of harms include the antiretroviral drug abacavir,[15] where HLA typing helps identify patients at high risk of abacavir toxicity. Thus a key part of stratified medicine research is to identify suitable tests for predicting treatment response from specific interventions.

The use of HER-2 status in breast cancer management illustrates how tests of differential treatment response are often thought of as binary factors: a biomarker is classed as positive or negative, or laboratory values are deemed low or high. Such dichotomisation facilitates clinical decision making and is used in most examples described in this paper. However, many tests have original values measured on an ordinal or a continuous scale. Similarly if prognostic models[10] are considered as tests, they usually produce a continuous risk score for each individual; the same applies to gene signatures or related indices derived from high dimensional data. Statistically, there is more power and less potential for bias if such tests are evaluated on their original scale (see later) rather than being dichotomised by means of a cut point[10]; categorisation may then be done after analysis to aid clinical strategies. For example, Flynn et al derived a prognostic model to identify patients with back pain who would respond well to manipulation rather than to other types of treatment such as exercise.[16] Some trials randomising patients to these treatments found that patients with positive scores from the model had greater relative and absolute benefits from manipulation than those with negative scores.[17 18]

Thus stratified medicine uses baseline information about a patient's likely response to treatment to tailor treatment decisions. This is different from stepped[19] or adaptive[20] models of care in which tailoring of treatment depends on the patient's actual response to previously offered treatment, with a sequence of interventions (which may differ in intensity, duration, cost, or complexity) being offered to those who have not responded sufficiently. Our focus here, though, is on the initial stratification of treatment based on the predicted (rather than actual) response to treatment.

Why is prognosis research important for stratified medicine?

Prognosis research is a fundamental component of stratified medicine because it contributes evidence at multiple stages in translation (see fig 1⇓ as an example). We now consider each of these stages in turn.

Assessing priorities for stratified medicine

Targeting interventions at defined patient strata is likely to be more important in some disease-treatment combinations than in others, and prognosis research can help prioritise areas for research. Several questions arise. First, is there clinically important variation in prognosis across individuals?[1] For example, among people with symptomatic severe aortic stenosis, one year survival is poor and valve replacement or implementation is the default option. By contrast, among people with aortic regurgitation, one year survival is better, and so tools to help decide when and for which patients valve replacement would yield the greatest benefit, and incur the least harm, would be a substantial advance. Second, is the intervention in question associated with a substantial risk of harm or cost? Third, for drug interventions, is there robust evidence of important individual variation in metabolism or pharmacological effect? For example, it has been claimed that some individuals have "clinical aspirin resistance" if they sustain a cardiovascular event despite aspirin prophylaxis. However, because of the lack of an optimal assay of platelet function and the paucity of high quality epidemiological data, it is unclear to what extent this observation reflects true pharmacological resistance to aspirin, non-adherence to medication,[21] the expected reduction but not abolition of cardiovascular risk from aspirin treatment, or some combination of these factors.

Discovery and candidate approaches to developing new tests

Prognosis research is important to identify which factors to study as potential predictors of differential treatment response, which might lead to a new prototype test (left hand of translational pathway in fig 1⇓). Prognostic factors, which were discussed in paper 2 of our series,[3] are characteristics associated with a particular outcome even in the absence of specific treatment. Prognostic factors with causal or mechanistically relevant effects are also potential predictors of differential treatment response. For example, among people with atrial fibrillation, age influences both response to warfarin and risk of stroke, and so is a both a prognostic factor[3] and a factor that predicts differential treatment response.

However, most prognostic factors do not also predict differential treatment response.[22] That is, they identify groups of patients with different absolute outcome risks, but not groups with different relative risks for a particular treatment. Conversely, a factor that predicts differential treatment response is not necessarily a prognostic factor. That is, some factors (such as those that influence the metabolism or elimination of a specific drug) may influence the response to treatment (that is, they modify relative risk) without affecting prognosis in the absence of treatment (that is, they do not change absolute risk). For example, the CYP2C9 and VKORC1 genotypes are associated with differential warfarin response but do not influence the risk of stroke in the absence of warfarin treatment.[23]

DNA based, genome-wide association studies (genomics) and mRNA based gene expression profiling (transcriptomics) of disease affected tissues are beginning to uncover new and, in some cases, unanticipated disease mechanisms and factors that potentially predict differential treatment response.

Evaluation in randomised trials

Once a factor potentially predicting differential treatment response has been identified the next step is to evaluate it, ideally as an a priori primary objective within a randomised trial of the specific therapy in question. Figure 2⇓ illustrates such a comparison of outcomes in treated and control groups, separately among factor positive and factor negative individuals. However, few individual trials are large enough to assess reliably whether a factor is truly predictive of treatment response as a primary objective, so evidence may often appear gradually, from secondary analyses of existing randomised trials and then their meta-analysis. This process was used for examining whether tamoxifen treatment of breast cancer differed according to the oestrogen receptor status of the breast cancer.[24]

Evaluations of factors that may predict differential treatment response become more pressing when a drug fails in late stage trials after substantial research investment; there is then intense interest in moving from targeting all people to identifying those specific patients who may benefit. For example, gefitinib in advanced non-small cell lung cancer failed to show a survival benefit among all patients, and this stimulated exploratory analyses in relation to epidermal growth factor receptor status.[14] Even in trials that do show a positive average effect of a drug, there may still be some patients who hardly benefit from the drug, and it is clearly important to identify this subgroup.[25] However, it is notoriously difficult to identify genuine predictors of differential treatment from single trials, as such investigations are usually exploratory with high potential for type I and type II errors (see later).

Assessment of tests as a health technology

Even seemingly robust evidence for the existence of a factor that predicts differential treatment response does not guarantee that it will be effective when used as a test in clinical practice to inform therapeutic decisions. Consider the example of pharmacogenetic testing to guide warfarin dosing. Here the testing, not the drug, is the technology being evaluated. In a high quality meta-analysis of nine observational studies (2775 patients),[26] CYP2C9*2 and CYP2C9*3 alleles were associated with a requirement for a lower warfarin dose and an increased risk of bleeding. Despite this clear association, which is unlikely to have arisen by chance, a systematic review of three randomised controlled trials did not provide evidence in favour of warfarin dosing based on genetic information in comparison with standard clinical care with respect to bleeding rate or time spent in the therapeutic range.[27]

Cost effectiveness evaluations

Decision analytic models are important for the evaluation of the cost effectiveness of stratified therapeutic strategies.[28-30] These models require valid estimates of prognosis under different scenarios, based on treatment with and without knowledge of the predictor of differential treatment response. Such models are important for policy makers because they evaluate strategies which are unlikely to be evaluated within trials. For example, decision analysis comparing different strategies for assessing HER-2 status to decide on treating breast cancer with trastuzumab found that fluorescent in situ hybridisation testing for all patients, with one year of adjuvant treatment with trastuzumab for those who were positive, was associated with the longest quality adjusted survival, with an estimated cost per quality adjusted life year gained of €41 500 (£32 600, $51 200).[31 32]

Healthcare policy and delivery

Health services research is required to examine variations in the uptake of using tests to predict differential treatment response,[33] the validity of these tests,[34] and variations in treatments based on test results. Prognosis research also examines endpoints in relation to these variations, allowing, for example, national estimates to be made of the number of endpoints averted by current levels of testing.[35]

Once incorporated in clinical practice guidelines[4] and usual clinical care, tests that predict differential treatment response may help define the disease and how it is characterised. This is termed "back translation." For example, in breast cancer, HER-2 and oestrogen receptor status are predictors of differential treatment response, and so their measurement is now integral to the definition of the disease upon diagnosis.

Premature implementation of stratified medicine approaches into clinical practice may be harmful if people who might otherwise benefit from treatment are denied access. For example, carriage of a variant of the KIF6 gene was associated with a higher risk of coronary heart disease events and a smaller reduction in event rate from statin treatment in a genetic substudy from a randomised trial.[36] It would have been premature to implement these findings; indeed, a later, larger meta-analysis of case-control studies of myocardial infarction casts doubt on the role of this variant in coronary heart disease and prediction of statin response,[37] arguing that statins should be used according to existing guidelines without any genetic testing

Recommendations for improving prognosis research for stratified medicine

Several methodological challenges and current research deficiencies need to be addressed in this field. Currently we lack a systematic framework for guiding research on stratified medicine, and standards must be raised. Many of the recommendations highlighted across the PROGRESS series (see supplementary table on bmj.com) are relevant. For example, integrated standards of design, analysis, and reporting should be developed across the stages of discovery, replication, and evaluation of factors that potentially predict differential treatment response[38-43] (recommendation 10 in supplementary table). Here we highlight four key areas, with recommendations for improvement.

False negative findings (type II errors)

There are important problems with statistical analyses, which should be addressed by having a statistical analysis plan in the protocol and by a greater appreciation of the potential for type I and II errors that may lead to inappropriate conclusions (recommendation 13 in supplementary table).

Most randomised trials are not designed with the statistical power to detect a factor truly predictive of differential treatment effect, should it exist, and so may wrongly conclude that a particular factor is not useful as a predictive test when actually it is.[44 45] To increase power and reduce the opportunity for false negatives, we recommend that meta-analyses based on individual participant data from multiple trials are facilitated (recommendation 17).[46] This approach was crucial in establishing the role of oestrogen receptor status for the targeting of tamoxifen treatment in breast cancer,[24] and researchers can support its greater use by initiating collaborative groups and data sharing.[46] Another cause of false negative findings is the aforementioned dichotomisation of continuous factors that may predict treatment response, which reduces power further. Statistical methods are available to screen a large number of continuous factors on their original scale and identify their potential interactions with treatment.[47] Identified interactions should be interpreted as hypothesis generating and replication sought in other studies, and meta-analyses. Results for all interactions and subgroups considered should be clearly reported regardless of their significance (recommendation 15), and guidelines for such reporting need development.

False positive findings (type I errors)

Subgroup analyses can provide valuable, albeit predominately exploratory information, about factors that potentially predict treatment response if they are performed in accordance with recommendations and guidelines[38 48] (recommendation 13). However, inappropriate subgroup analysis of trial data can give spurious evidence for stratified medicine. Firstly, because of the large number of potential factors to consider, appropriate correction for multiple statistical testing is required to reduce the risk of false positives arising by chance.[45 49] Alternatively, we recommend that such analyses should be recognised as exploratory and require replication using new data from related studies and in meta-analysis of individual participant data (recommendations 17 and 9).

Secondly, the choice and handling of endpoints can influence interpretation of evidence about whether a factor predicts treatment response. For example, in a field synopsis of pharmacogenetic studies, there was evidence of bias in which positive findings were more likely when examining surrogate markers of treatment effects rather than the more clinically relevant endpoints such as a disease complication or death.[50]

Thirdly, arbitrary or "data dredging" categorisation of continuous factors and continuous outcomes can easily bias findings toward a significant result, particularly if analyses are repeated for multiple cut-offs until a categorisation is found that provides the most significant P value.[51] Continuous factors should rather be analysed on their continuous scale to avoid this.

Fourthly, as Senn has argued, studies claiming to distinguish responders (say 70% of people) and non-responders (30%) after a single exposure to a drug are also consistent with an alternative explanation that 100% of patients respond 70% of the time, which would indicate the absence of differential response to treatment.[52]

Fifthly, a meta-analysis of summary data from trials may also give misleading positive results, and a meta-analysis of individual participant data is preferred. For example, fig 3⇓ shows a meta-analysis of summary data from 10 trials suggesting women experience a greater and clinically important reduction in blood pressure from hypertension treatment than men. By contrast, in a meta-analysis of individual participant data from the same trials this apparent sex-treatment interaction was found to be small and not clinically important.[46 53] The discrepant findings were caused by study level confounding when looking at aggregated relationships across trials, rather than investigating patient level relationships within trials using individual participant data.

Away from trials, many consider molecular and microarray data are the key to stratified medicine, but so far the high expectations have not been met, and a more realistic view is important.[3] The large number of variables collected in a relatively small number of patients results in severe methodological problems,[3] and type I errors are again a particular concern.

Analyses restricted to just individuals testing positive for a factor, or just individuals receiving treatment

Robust trial designs to identify factors that truly predict differential treatment response should ideally involve the four groups of patients illustrated in the lower panel of fig 2⇓ so that the difference in treatment effect between patients who are positive for the factor and those who are negative can be estimated (recommendation 22). However, such a design is often not carried out.

Increasingly, drug trials are being undertaken exclusively among individuals who test positive for a potential (but unproved) factor that predicts differential treatment response (upper panel of fig 4⇓). For example, a randomised trial of heart rate lowering drug ivabradine failed to show a benefit in primary outcome of events among people with stable coronary disease, but subgroup analysis suggested a benefit among those with higher heart rates.[3] The subsequent trial was confined to people with higher heart rates.

Emerging trial designs even propose the integration of drug evaluation with the discovery and evaluation of novel biomarker signatures in real time.[3-55] Such studies are sometimes referred to as enrichment trials because, by selecting people in whom the treatment effect is hypothesised to be large, they provide a mechanism for reducing the sample size of a trial. This is only a sensible approach as long as inferences are restricted to the selected patients in the trial. In particular, such trials cannot then compare outcomes between patients with positive and negative factor values, and so cannot assess whether the relative treatment effect (or differences in absolute risk) are indeed smaller in individuals with negative values for the factor, let alone the differences in absolute risk.

Of much more concern are observational analyses (either within or outside the framework of a trial) confined to just those who are treated, as then no comparison can be made with control patients and thus the treatment effect cannot be estimated. In this type of approach, to be able to conclude that a factor truly predicts treatment response, one must assume that the factor does not influence the outcome of interest in the absence of treatment (lower panel of fig 4⇓). If the factor is associated with outcome in both treated and untreated individuals, then it may be a prognostic factor (as discussed in paper 2 of our series[3]) but not predictive of treatment response. Thus, the approach is

more correctly interpreted as an evaluation of a prognostic factor among those who are treated, but this is often not recognised.

Biological reasoning and prioritisation of funding areas

Statistical evidence of an interaction between a particular factor and treatment response should ideally be explained by biological reasoning and by understanding the mechanism by which response is modified. For instance, for drug interventions, clinicians and policy makers are more likely to believe that a factor truly modifies treatment response if there is a well reasoned biological mechanism in addition to statistical significance. Indeed, stratified medicine research may be entirely motivated by such a biological mechanism in the first place, and funders should prioritise stratified medicine investigations that have such plausibility. "Biological mechanism" should be interpreted in a broad sense here, since behavioural and sociocultural factors may be of equal importance (and have plausible mechanisms for their effects on health outcomes) to biologically measured factors and pathways.

There should be rigorous evaluation of the impact of "personalised medicine" approaches on health outcomes, including comparisons of approaches based on targeting intervention (with prognostic models or factors that predict differential treatment response) and "all comer" approaches (recommendation 23 in supplementary table). In certain situations subgroups with weaker treatment effects on relative risk may have the greater potential benefit in terms of absolute risk. Uncertainty about treatment effects is usually greater in low risk groups, and adequately powered prognosis research is required.

Funders and policy makers should also recognise that a treatment may benefit all patients even when there is a factor that predicts treatment response. In this situation, patients testing negative for the factor will still benefit from the treatment, and so treatment policies should not automatically exclude such patients.

Industry interest (drug, device, biomarker, information technology) in prognosis research including tests for stratified medicine (sometimes called "companion diagnostics"), drug safety, outcomes research and real world evidence is growing. Appropriate models of industry and publicly funded prognosis research should be developed which allow unbiased inference. (recommendation 24 in supplementary table).

Conclusions

In this article we have illustrated and described how prognosis research contributes important evidence in discovering, developing, evaluating, and implementing new approaches in stratified medicine, especially in identifying factors that truly predict differential treatment response. Such research faces many challenges, and often current study designs and statistical analyses are substandard. We have provided recommendations with the aim of accelerating the potential of prognosis research in this context, and these build on others presented throughout our PROGRESS series to improve the care, treatment, and clinical outcomes of individual patients.

We thank John Scadding, emeritus dean at the Royal Society of Medicine for his support of the PROGRESS Group, contributions to discussions, and helpful comments on drafts of the manuscripts. We thank Ruzan Udumyan for assistance in drawing figures and preparing manuscripts. We thank Virginia Barbour and Trish Groves for contributing to the workshops and their support for the series. We thank Lucy Chappell (King's College London) for her valuable help as guest editor on the PROGRESS series.

Contributors: HH, RDR, SS, and DGA initiated the PROGRESS Group, organised the three workshops, coordinated the writing groups, and were the scientific writing editors for all the papers in the PROGRESS series. HH, RDR, and DGA are the guarantors for this paper.

All members of PROGRESS Group contributed through workshops and discussions to the development of the article series. AH wrote the first version, and the other authors contributed to the conception of the paper and to revising earlier versions, and approved the final version. RDR revised the document in light of coauthor comments and restructured it to align with the other PROGRESS papers. All members of PROGRESS Group contributed through workshops and discussions to the development of the article series.

Members of the PROGRESS Group: Keith Abrams (UK), Doug Altman (UK), Andrew Briggs (UK), Nils Brunner (Denmark), Peter Croft (UK), Jill Hayden (Canada), Aroon D Hingorani (UK), Harry Hemingway (UK), Panayiotis Kyzas (UK), Núria Malats (Spain), Karel Moons (Netherlands), George Peat (UK), Pablo Perel (UK), Richard Riley (UK), Ian Roberts (UK), Willi Sauerbrei (Germany), Sara Schroter (UK), Ewout Steyerberg (Netherlands), Adam Timmis (UK), Daniëlle van der Windt (UK).

Funding: This series had no explicit funding, but some of the authors were supported by research grants: PROGRESS is supported by a Partnership grant from the Medical Research Council (G0902393), involving University College London (HH, AH), Oxford University (DGA), Birmingham University (RDR), London School of Hygiene and Tropical Medicine (I Roberts, P Perel), Keele University (P Croft, DAvdW), and Queen Mary University London (A Timmis). DGA is supported by a programme grant from Cancer Research UK (C5529). HH is supported by grants from the UK National Institute for Health Research (RP-PG-0407-10314) and the Wellcome Trust (086091/Z/08/Z). The work of HH and ADH is supported by the Health eResearch Centre Network (HERC-UK), funded by the Medical Research Council in partnership with Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the National Institute of Health Research, the National Institute for Social Care and Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates) and the Wellcome Trust (the views expressed in this paper are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health). EWS was supported by The Netherlands Organization for Scientific Research (grant 9120.8004) and the Center for Translational Molecular Medicine (PCMM project, grant 03O-203). RDR is supported by the MRC Midlands Hub for Trials Methodology Research (MRC Grant G0800808). DAvdW is supported by the Arthritis Research UK Centre of Excellence in Primary Care.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work. SS is a full time employee of the BMJ Group but is not involved in deciding which manuscripts are accepted for publication.

Provenance and peer review: Not commissioned; externally peer reviewed. In order to disseminate the output widely, these papers are being published jointly between *BMJ* (PROGRESS papers 1 and 4) and *PLoS Medicine* (PROGRESS papers 2 and 3). As one of the authors is a member of staff at BMJ Group, the handling editor at the BMJ for the manuscripts was an external guest editor, Lucy Chappell (King's College London).

Summary points

The PROGRESS series (www.progress-partnership.org) sets out a framework of four interlinked prognosis research themes and provides examples from several disease fields to show why evidence from prognosis research is crucial to inform all points in the translation of biomedical and health related research into better patient outcomes. Recommendations are made in each of the four papers to improve current research standards

What is prognosis research? Prognosis research seeks to understand and improve future outcomes in people with a given disease or health condition. However, there is increasing evidence that prognosis research standards need to be improved

Why is prognosis research important? More people now live with disease and conditions that impair health than at any other time in history; prognosis research provides crucial evidence for translating findings from the laboratory to humans, and from clinical research to clinical practice

Stratified medicine involves tailoring therapeutic decisions for specific, often biologically distinct, individuals, the aim being to maximise benefit and reduce harm from treatment, or to rescue a treatment that fails to show overall benefit in unselected patients but does benefit specific patients

Stratified medicine can use absolute risks. When a treatment effect measured on a relative scale (such as relative risk) is the same for all patients, those with the highest absolute risk will derive the largest absolute benefit from the treatment

When the relative treatment effect is inconsistent across patients, stratified medicine can use tests which measure factors (such as biomarker levels or genotypes) that predict individual treatment response. However, the clinical use of such tests is currently small, and rigorous evidence of impact is sometimes lacking, with flaws in study design, analysis, and reporting leading to potentially spurious evidence either for or against a factor

Research to identify factors that truly predict treatment effect could be improved by:
Labelling exploratory analyses as exploratory, to minimise false positive findings

Increasing statistical power by designing trials with adequate sample sizes, facilitating collaborations across research groups and meta-analyses of individual participant data from multiple trials, and by analysing continuous factors on their original scale

Estimating, for a truly binary factor, the difference in relative treatment effect between positive and negative groups within randomised trials that include both factor positive and factor negative patients in both control and treatment groups

Considering biological or other mechanisms for modification of treatment response, either to motivate new research or to support statistical evidence that a factor interacts with treatment

Prognosis research in general should play a more central role in stratified medicine research: from identifying conditions with clinically important differences in absolute risk of outcome across patients, to identifying factors that predict individual treatment response, and to examining the cost and impact of implementing stratified medicine approaches in practice

The other papers in the series are:
PROGRESS 1: *BMJ* 2013, doi:10.1136/bmj.e5595

PROGRESS 2: *PLoS Med* 2013, doi:10.1371.journal/pmed.1001380

PROGRESS 3: *PLoS Med* 2013, doi:10.1371.journal/pmed.1001381

1 Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ* 2013;346:e5595.
2 Hudis CA. Trastuzumab—mechanism of action and use in clinical practice. *N Engl J Med* 2007;357:39-51.
3 Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis research strategy (PROGRESS) 2: Prognostic factor research. *PLoS Med* 2013, doi:10.1371.journal/pmed.1001380.
4 Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J Clin Oncol* 2007;25:118-45.
5 Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 1987;235:177-82.
6 Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med* 2010;363:301-4.
7 House of Lords Science and Technology Committee. *Genomic medicine. Volume II: evidence* . Stationery Office, 2009. www.publications.parliament.uk/pa/ld200809/ldselect/ldsctech/107/107ii.pdf
8 Trusheim MR, Berndt ER, Douglas FL. Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nat Rev Drug Discov* 2007;6:287-93.
9 Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. *Lancet* 2005;365:256-65.
10 Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013, doi:10.1371.journal/pmed.1001381.
11 Hingorani AD, Hemingway H. How should we balance individual and population benefits of statins for preventing cardiovascular disease? *BMJ* 2011;342:c6244.
12 Hill JC, Whitehurst DG, Lewis M, Bryan S, Dunn KM, Foster NE, et al. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet* 2011;378:1560-71.
13 Capdeville R, Buchdunger E, Zimmermann J, Matter A. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nat Rev Drug Discov* 2002;1:493-502.
14 Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* 2009;361:947-57.
15 Mallal S, Nolan D, Witt C, Masel G, Martin AM, Moore C, et al. Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* 2002;359:727-32.
16 Flynn MR, Barrett C, Cosio FG, Gitt AK, Wallentin L, Kearney P, et al. The Cardiology Audit and Registration Data Standards (CARDS), European data standards for clinical cardiology practice. *Eur Heart J* 2005;26:308-13.
17 Brennan GP, Fritz JM, Hunter SJ, Thackeray A, Delitto A, Erhard RE. Identifying subgroups of patients with acute/subacute “nonspecific” low back pain: results of a randomized clinical trial. *Spine* 2006;31:623-31.
18 Childs JD, Fritz JM, Flynn TW, Irrgang JJ, Johnson KK, Majkowski GR, et al. A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: a validation study. *Ann Intern Med* 2004;141:920-8.
19 Von KM, Moore JC. Stepped care for back pain: activating approaches for primary care. *Ann Intern Med* 2001;134:911-7.
20 Almirall D, Compton SN, Gunlicks-Stoessel M, Duan N, Murphy SA. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Stat Med* 2012;31:1887-902.
21 Hankey GJ, Eikelboom JW. Aspirin resistance. *Lancet* 2006;367:606-17.
22 Clark GM. Prognostic factors versus predictive factors: examples from a clinical trial of erlotinib. *Mol Oncol* 2008;1:406-12.
23 Klein TE, Altman RB, Eriksson N, Gage BF, Kimmel SE, Lee MT, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med* 2009;360:753-64.
24 Early Breast Cancer Trialists’ Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet* 1998;351:1451-67.
25 Royston P, Sauerbrei W, Ritchie A. Is treatment with interferon-alpha effective in all patients with metastatic renal carcinoma? A new approach to the investigation of interactions. *Br J Cancer* 2004;90:794-9.
26 Sanderson S, Emery J, Higgins J. CYP2C9 gene variants, drug dose, and bleeding risk in warfarin-treated patients: a HuGEnet systematic review and meta-analysis. *Genet Med* 2005;7:97-104.
27 Kangelaris KN, Bent S, Nussbaum RL, Garcia DA, Tice JA. Genetic testing before anticoagulation? A systematic review of pharmacogenetic dosing of warfarin. *J Gen Intern Med* 2009;24:656-64.
28 Henderson R, Keiding N. Individual survival time prediction using statistical models. *J Med Ethics* 2005;31:703-6.
29 Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MSV, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation* 2009;119:2408-16.
30 Moons KG. Criteria for scientific evaluation of novel markers: a perspective. *Clin Chem* 2010;56:537-41.
31 Lidgren M, Jonsson B, Rehnberg C, Willking N, Bergh J. Cost-effectiveness of HER2 testing and 1-year adjuvant trastuzumab therapy for early breast cancer. *Ann Oncol* 2008;19:487-95.
32 Williams AH, Cookson RA. Equity-efficiency trade-offs in health technology assessment. *Int J Technol Assess Health Care* 2006;22:1-9.
33 Woelderink A, Ibarreta D, Hopkins MM, Rodriguez-Cerezo E. The current clinical practice of pharmacogenetic testing in Europe: TPMT and HER2 as case studies. *Pharmacogenomics J* 2006;6:3-7.
34 Nakhleh RE, Grimm EE, Idowu MO, Souers RJ, Fitzgibbons PL. Laboratory compliance with the American Society of Clinical Oncology/College of American Pathologists guidelines for human epidermal growth factor receptor 2 testing: a College of American Pathologists survey of 757 laboratories. *Arch Pathol Lab Med* 2010;134:728-34.
35 Danese MD, Lalla D, Brammer M, Doan Q, Knopf K. Estimating recurrences prevented from using trastuzumab in HER-2/neu-positive adjuvant breast cancer in the United States. *Cancer* 2010 15;116:5575-83.

36 Li Y, Iakoubova OA, Shiffman D, Devlin JJ, Forrester JS, Superko HR. KIF6 polymorphism as a predictor of risk of coronary events and of clinical event reduction by statin therapy. *Am J Cardiol* 2010;106:994-8.
37 Assimes TL, Holm H, Kathiresan S, Reilly MP, Thorleifsson G, Voight BF, et al. Lack of association between the Trp719Arg polymorphism in kinesin-like protein-6 and coronary artery disease in 19 case-control studies. *J Am Coll Cardiol* 2010;56:1552-63.
38 Fayers PM, King MT. How to guarantee finding a statistically significant difference: the use and abuse of subgroup analyses. *Qual Life Res* 2009;18:527-30.
39 Guillemin F. Primer: the fallacy of subgroup analysis. *Nat Clin Pract Rheumatol* 2007;3:407-13.
40 Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med* 2007;26:5512-28.
41 Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010;340:c117.
42 Willett WC. The search for truth must go beyond statistics. *Epidemiology* 2008;19:655-6.
43 Sauerbrei W, Royston P. Modelling to extract more information from clinical trials data: on some roles for the bootstrap. *Stat Med* 2007;26:4989-5001.
44 Guyatt G, Rennie D, Meade M, Cook D. When to believe a subgroup analysis. In: *Users' guides to the medical literature: a manual for evidence-based clinical practice* . 2nd ed. AMA Press, 2008.
45 Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;266:93-8.
46 Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;340:c221.
47 Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Stat Med* 2004;23:2509-25.
48 Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064-9.
49 Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ* 2003;326:219.
50 Holmes MV, Shah T, Vickery C, Smeeth L, Hingorani AD, Casas JP. Fulfilling the promise of personalized medicine? Systematic review and field synopsis of pharmacogenetic studies. *PLoS One* 2009;4:e7960.
51 Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;25:127-41.
52 Senn S. Individual response to treatment: is it a valid assumption? *BMJ* 2004;329:966-8.
53 Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Stat Med* 2008;27:1870-93.
54 Mandrekar JN, Mandrekar SJ. Case-control study design: what, when, and why? *J Thorac Oncol* 2008;3:1371-2.
55 Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: lessons from real trials. *Clin Trials* 2010;7:567-73.

Accepted: 29 July 2012

Cite this as: *BMJ* 2013;346:e5793

Figures

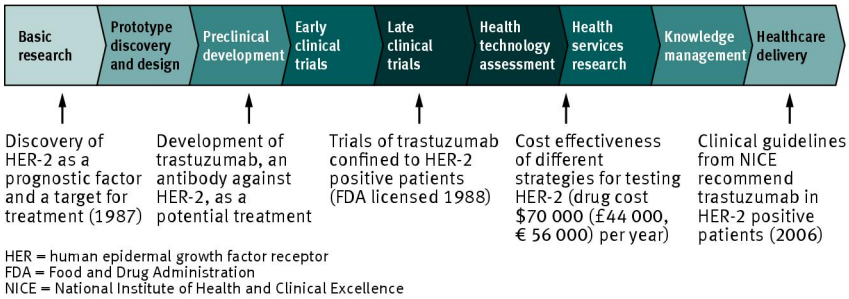

Fig 1 Example of stratified medicines research, with translation from discovery of human epidermal growth factor receptor 2 (HER-2) status as a prognostic factor for metastatic breast cancer[5] to development of trastuzumab treatment and use in clinical practice. Path element adapted from chart 7.1 in the Cooksey report (2006) (made available for use through the Open Government License)

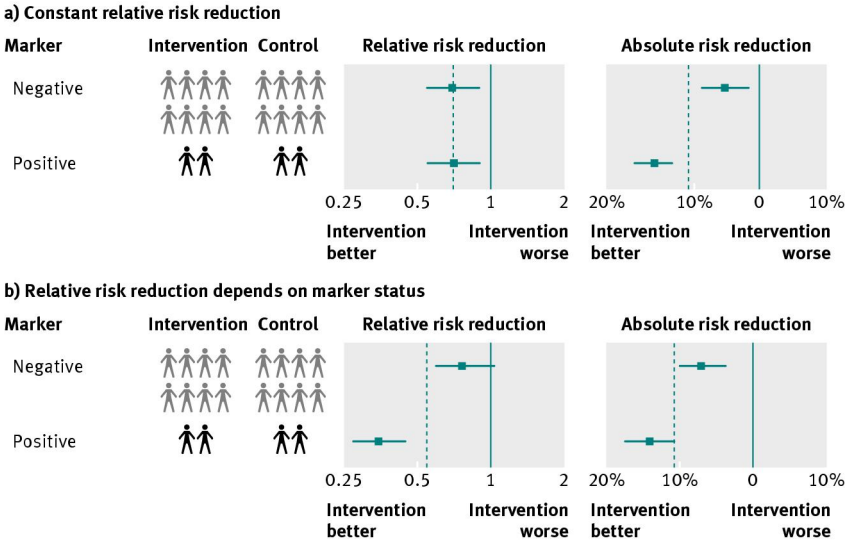

Fig 2 Estimated treatment effect in subgroups defined according to (upper panel) risk from a prognostic model and (lower panel) a factor that predicts differential treatment response. The prevalence of positive factor and high risk is shown, arbitrarily, as 20%. The dotted vertical line shows the overall treatment effect, the centre of each box shows the effect estimate, and the horizontal lines show confidence intervals

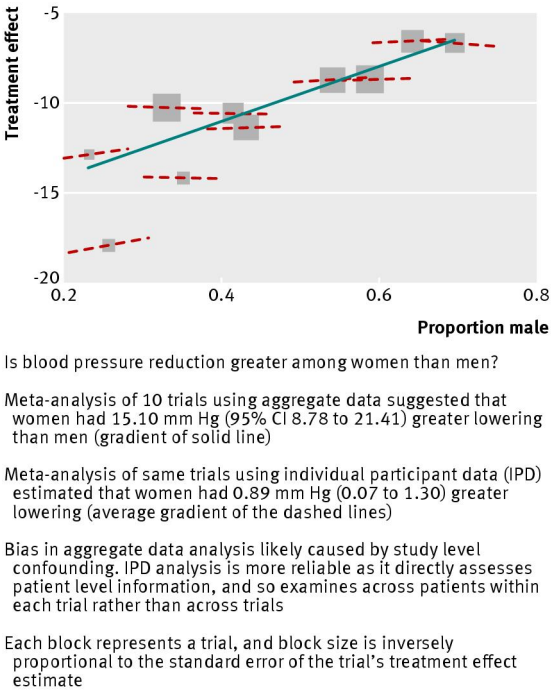

Fig 3 Example of spurious finding in meta-analysis of summary data refuted by meta-analysis of individual participant data: whether antihypertensive treatment has a greater effect in women than men (reproduced with permission from Riley et al[46 53])

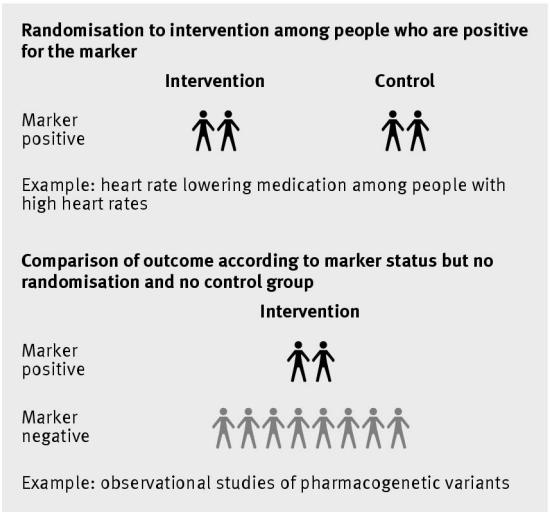

Fig 4 Commonly used (but suboptimal) study designs in assessment of a factor that potentially predicts differential treatment response