PRECISION MEDICINE

The evidence landscape in precision medicine

Spencer Phillips Hey[1,2]*, Cory V. Gerlach[3,4,5], Garrett Dunlap[3,6], Vinay Prasad[7], Aaron S. Kesselheim[1,2]

Precision medicine is beginning to make an impact on the treatment of different diseases, but there are still challenges that must be overcome, such as the complexity of interventions, the need for marker validation, and the level of evidence necessary to demonstrate effectiveness. In this Perspective, we describe how evidence landscapes can help to address these challenges.

INTRODUCTION

Precision medicine is an increasingly important prospect for patients and physicians alike. New insights into the underlying biology of cancer have made it possible to prospectively test patients for particular genetic mutations that drive tumor growth. Individuals with cancer carrying a particular genetic mutation can be steered toward therapies that directly target that pathogenic pathway while sparing patients without the mutation from the unnecessary toxicity and burden of an ineffective intervention. Yet, despite the evident promise of this therapeutic strategy, there are several challenges that must be overcome to successfully translate such targeted therapies from the research setting to the clinic. For example, one challenge arises from the fact that the unit of testing in precision medicine is a complex intervention ensemble that combines a therapeutic agent, a marker, and a diagnostic assay for detecting that marker (*1*). Rigorous testing of this complex intervention ensemble requires that each component—treatment, marker, and assay—has been optimized for a given condition.

A second challenge arises from the fact that diagnostic assays require their own multistep development and validation process, which involves assessment of the assay's preanalytical validity (proper specimen handling and processing), analytical validity (test accuracy, reliability, and reproducibility), clinical validity (strong association between test result and a clinical outcome of interest), and clinical utility (use of the test to direct patient care results in a more favorable risk-benefit balance than nonuse of the test) (*2*). Failure to complete each of these steps can result in poor uptake of the diagnostic technology, patient misclassification, worse patient outcomes, or unnecessary health care spending (*3*, *4*).

A third challenge arises from disagreement about the level of evidence necessary to demonstrate the effectiveness of the intervention ensemble. Simon *et al.*'s influential evidence hierarchy for precision medicine marker validation (*5*) stipulates that level 1A evidence requires randomized controlled trials (RCTs) that stratify patients according to their marker status and then randomize them to a therapy, thus providing a rigorous test of the interaction between a patient's marker test result and their clinical outcomes with a particular intervention. Level 1B evidence can be produced by retrospective analyses of RCTs that analyze the interaction between marker status and treatment response after the RCT has been carried out (*5*). However, a recent analysis of targeted drugs approved by the U.S. Food and Drug Administration (FDA) found that the majority were approved on the basis of evidence from enrichment trials only—that is, trials that restricted enrollment to patients who tested positive for the marker of interest (*6*). Whereas enrichment trials can provide evidence about how a therapy works in the marker-positive patient population, their exclusion of the marker-negative population means that they do not provide evidence about the clinical validity of the marker (*7*). Thus, for most approved targeted medicines, the complete intervention ensemble has not been fully tested, and it is not known whether the marker diagnostic is actually a necessary component of the therapeutic strategy.

We believe that overcoming these challenges for precision medicine requires new approaches to scientific coordination and communication (*1*). One approach could be to have large research funders and institutions collaborate to prospectively map out the parameter space for intervention ensembles—that is, creating an evidence landscape—thereby providing a public platform to help track the accumulation of evidence and identify promising avenues for further research. This model would complement efforts for improving reporting (*8*) and preregistration (*9*) of intervention ensemble experiments. It would also help physicians and other knowledge users to better assess the state of supporting evidence for marker-based treatment strategies.

Here, we illustrate how such an evidence landscape for precision medicine could be constructed using as an example the testing of mutations in v-Raf murine sarcoma viral oncogene homolog B (BRAF) in metastatic melanomas. Testing for BRAF mutations is widely used to select individuals with metastatic melanoma for treatment with one of the two approved BRAF inhibitors—vemurafenib, approved by the FDA in 2011, and dabrafenib, approved in 2013. Both of these drugs target BRAF V600 mutations, and both were approved with companion diagnostic assays, the Cobas 4800 BRAF V600 mutation test (for vemurafenib) and the THxID BRAF test (for dabrafenib).

Companion diagnostics are described by the FDA as "essential for the safe and effective use of the corresponding therapeutic product" (*10*). FDA approval of the BRAF inhibitors and the companion diagnostic tests was based on enrichment trials that only enrolled patients who had tested positive for BRAF V600 mutations (*11*, *12*). Thus, in addition to illustrating the power of constructing an evidence landscape for precision medicine, this example is also an opportunity to critically examine what is known (and remains

[1]Harvard Center for Bioethics, Harvard Medical School, Boston, MA, USA. [2]Program on Regulation, Therapeutics, and Law (PORTAL), Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [3]Harvard Program in Therapeutic Sciences, Harvard Medical School, Boston, MA, USA. [4]Department of Medicine, Renal Division, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [5]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [6]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, and Therapeutics Graduate Program, Harvard Medical School, Boston, MA, USA. [7]Division of Hematology and Medical Oncology, Knight Cancer Institute, Department of Preventive Medicine and Public Health, and Center for Health Care Ethics, Oregon Health and Science University, Portland, OR, USA.
*Corresponding author. Email: spencer_hey@hms.harvard.edu

unknown) about the clinical validity of an approved targeted treatment strategy.

THE PRECISION MEDICINE EVIDENCE LANDSCAPE

Given that typically there will be many different possible ensembles for any given drug/marker/disease combination and that there may be no a priori way to identify which ensembles are likely to be the most useful, achieving the goal of precision medicine requires a systematic and thorough exploration of the possibilities. We can therefore think of research and development of precision medicines as akin to a search through a multidimensional parameter space, or "evidence landscape," with the goal of identifying the sets of parameters that make for a clinically valid and useful ensemble.

One application of this evidence mapping for BRAF mutations in metastatic melanoma is depicted in Fig. 1 (see the interactive version at www.aerodatalab.org/braf). The Supplementary Materials provide a description of the search, data extraction, and visualization methods. For this evidence map, we held the BRAF mutations and disease constant and then defined the ensemble landscape in terms of three other parameters: therapeutic class, line of therapy, and assay method. Each vertex in this graph represents a possible ensemble—one combination of therapy, marker, assay method, and patient population (Fig. 1). If a particular ensemble has been studied and the results reported in the published literature, then we added a node at the corresponding vertex. The size of the node corresponds to the sample size of the published experiment; the color of the node represents the result of the experiment with respect to clinical validity. If the clinical study found a statistically significant association between a patient's BRAF mutation status and either objective response rate, progression-free survival, or overall survival, then the node was colored blue. If no significant associations between marker status and outcome were observed, then the node was colored orange. Whereas outcome representations can potentially be made more complex to capture additional details about each study's result, we believe this dichotomization is still useful as a first step to reveal the "hits" and "misses" in the landscape.

Before turning to discuss what this landscape shows us about BRAF testing in particular, we should first emphasize some of the general insights about intervention ensemble development revealed by this method.

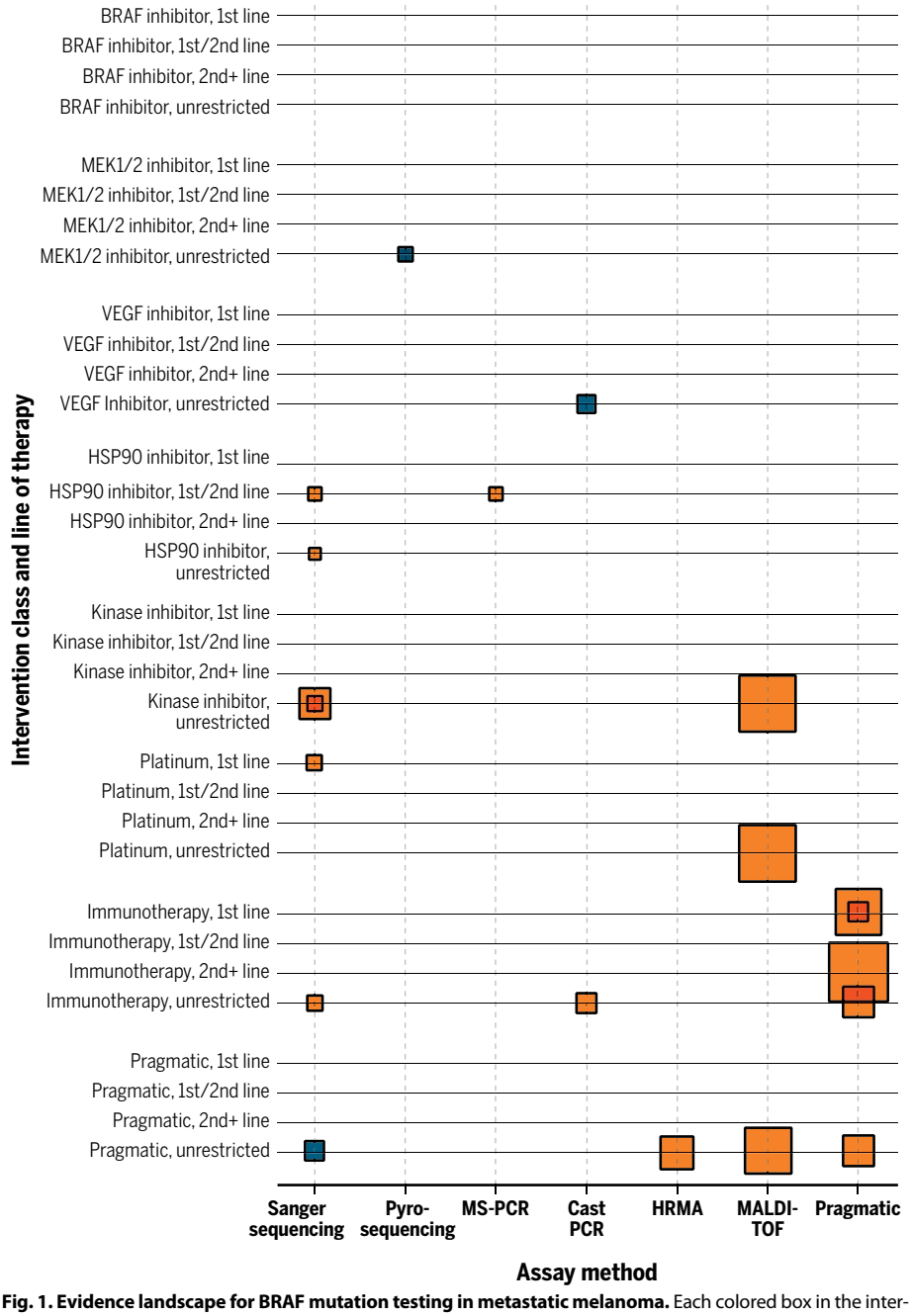

Fig. 1. Evidence landscape for BRAF mutation testing in metastatic melanoma. Each colored box in the intervention ensemble corresponds to an experimental evaluation of a particular ensemble, defined as the combination of therapeutic mechanism, line of therapy, and assay method. Box size corresponds to sample size (the number of patients/specimens evaluated). Box color corresponds to the association between BRAF mutation status and treatment response. If a statistically significant association was observed between either objective response rate, progression-free survival, or overall survival, then the box was colored blue; otherwise, the box was colored orange. Studies that did not restrict enrollment or inclusion to individuals receiving a particular type of therapy or that did not report the assay method used to determine BRAF mutation status are classified as pragmatic. An interactive version of this evidence landscape is available at www.aerodatalab.org/braf and can be accessed by the blue box surrounding the figure. The interactive map allows the user to drag a mouse over each node to see more information about the study or click on a node that links directly to the study's publication. This interactive implementation of the evidence landscape could be combined with a database for study registration and outcome reporting, thereby providing a useful platform to track the state of evidence and coordinate precision medicine research. MS-PCR, mutation-specific polymerase chain reaction; HRMA, high-resolution melting analysis; castPCR, competitive allele-specific TaqMan polymerase chain reaction; MALDI-TOF, matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry; VEGF, vascular endothelial growth factor; MEK1/2, MAPK kinase 1/2.

First, this was a highly simplified parameter space. The set of parameters constituting a complete ensemble includes many other elements, such as tissue sources (e.g., primary or metastatic or both), specimen storage method (e.g., paraffin embedded or frozen), and scoring or cutoff rules to divide the assay results into "marker-positive" or "marker-negative." Nevertheless, even for this simplified set of parameters, it becomes clear that the total parameter space for an ensemble is vast. Finding the optimal vertex (or "peak") of clinical validity in this landscape is therefore likely to require a systematic and coordinated research effort. This effort can only begin in earnest once investigators have adequately recognized the scope of the challenge, and we believe this method can help to make that scope explicit.

The evidence landscape also provides a comprehensive summary of the existing evidence surrounding a given ensemble. It shows which ensembles have been tested, what results have been replicated, which of the tested ensembles seem promising, and if there are regions of the space (i.e., particular parameters) that have been exhausted or overlooked. This is all valuable information when making decisions about the reliability of particular ensembles or when planning future investigations.

THE KNOWNS AND UNKNOWNS OF BRAF MUTATION TESTING

Turning now to our case study of BRAF mutation testing in metastatic melanoma, we believe that the pattern of research activity highlights several issues. First, few parameter combinations appear to have been formally explored and published. This may be surprising, given that BRAF mutation testing is now considered the standard practice for metastatic melanoma. However, the sparseness of exploration may be explained by the sheer number of experiments that would be required to thoroughly test all of the possible ensembles, as well as the limited financial incentives for cancer diagnostic development (*3*). This approach thus helps to sharpen questions about how to optimally (or strategically) explore the evidence landscape and most efficiently validate a precision medicine, given that there will never be sufficient research resources to explore it all.

This relates directly to a second finding. There have been no tests of the clinical validity of BRAF mutation testing for BRAF inhibitor therapies—the setting in which BRAF mutation testing is widely used to inform therapeutic decision-making. Given that this is a highly promising region of the ensemble landscape, it is unexpected to find no rigorous trials demonstrating the clinically meaningful interaction between the BRAF mutation status of individuals with melanoma and clinical outcomes with the BRAF inhibitors vemurafenib and dabrafenib. One explanation is that it is believed to be unethical to enroll patients without a BRAF V600 mutation in clinical trials of these BRAF inhibitors. This rationale was based on preclinical and early-phase evidence, suggesting that only individuals with melanomas carrying a BRAF mutation were likely to respond (*13*, *14*). Thus, the lack of studies in this region of the landscape (combined with limitations in research resources more generally) may be explained by a need to focus research efforts on translating the most promising interventions as quickly as possible—even if that entails uncertainty about the clinical validity of BRAF mutation testing.

Notably, however, at least some patients classified as having wild-type BRAF may still benefit from therapies that target BRAF mutations. One retrospective analysis comparing the results of multiple assays and multiple tissue specimens from 74 individuals with melanoma found discordant BRAF results for 14% of patients who were classified as having mutant BRAF by one test but wild-type BRAF by another (*15*). These patients included two individuals who had been classified as having wild-type BRAF but who were nevertheless treated with BRAF inhibitor therapy and benefited (*15*). Given the value—to clinicians, patients, and payers—of determining the clinical validity of a BRAF ensemble, we believe that it is still worth thinking critically about the circumstances under which preclinical evidence and a plausible biological rationale should lead the scientific community to skip over rigorous trials in the highly promising regions of an ensemble landscape.

Third, the pattern of exploration does not appear systematic, although that may be an artifact of how we have chosen to organize the parameters. Sanger sequencing is the most thoroughly explored assay method, used across a variety of drug classes and settings, but there are only a handful of published experiments examining the clinical validity of the other methods. For example, the two FDA-approved companion diagnostic tests for the BRAF inhibitors are both polymerase chain reaction assays, a method for which the clinical validity has only been evaluated in three studies. Most of the positive results in the BRAF ensemble landscape have not been replicated or followed up by confirmatory trials.

Finally, we created the category of "pragmatic" assays for our landscape to classify published studies that reported the association between BRAF mutation status and patient outcomes but did not report the assay method used to determine the BRAF mutation status of the individuals with melanoma (Fig. 1). In general, assay methods should not be assumed interchangeable, and not reporting the assay method undermines the interpretability and scientific value of the study (*16*). The choice of assay appears to matter for detecting BRAF mutations (*17*). For example, the approved companion diagnostic assay for vemurafenib, the Cobas V600 mutation test, has been shown to be less sensitive for detecting the V600K variant and thus was systematically misclassifying patients who may have benefited from the BRAF inhibitor therapy (*18*). Given that different clinics may be using different assays to determine the BRAF mutation status of their patients, reporting the particular method used in a clinical trial and understanding the relative error rates of different methods are vital for the safe and effective use of targeted therapies. Knowing the misclassification rates is also important for showing respect for individuals with melanoma who may be making treatment decisions on the basis of the assay results (*19*).

OPPORTUNITIES AND LIMITATIONS

As many commentators have noted, precision medicine's success depends upon the rigorous evaluation of multiple technologies (*2*). Funders and investigators could benefit by mapping out the important parameters of an intervention ensemble. Regulatory agencies could similarly use evidence landscapes to support their decisions when reviewing precision medicines and to inform their communications with drug developers, for example, encouraging drug developers to explore particular regions of the parameter space before submitting their product for approval.

There are also some important limitations to this approach. For example, the BRAF V600 mutant melanoma space may be one of the simpler ensembles, but the process of systematically reviewing the literature and assembling the evidence landscape still required considerable time and effort. It will require continual effort to keep the landscape up to date as new evidence emerges. However, we think this limitation could be overcome if larger research institutions (or a team of well-funded investigators) created a centralized platform for hosting evidence landscapes online.

Then, once an initial evidence review for an ensemble is completed, the process of keeping the evidence up to date could be distributed across the experts and investigators in the field who are already tracking the evidence.

Some of the properties of evidence landscapes, such as color-coding and limited numbers of axes, are also limitations of the approach. For example, although most of the studies in the BRAF ensemble landscape in Fig. 1 are orange (indicating no statistically significant associations with study outcomes), many of these studies were small and not sufficiently powered to detect the interaction between the therapy and the mutation status of individuals with melanoma. Therefore, this landscape should not be understood as showing that BRAF mutation testing is not useful, just that these studies do not show its utility. Similarly, there are other dimensions of the ensemble, such as assay cut-off points or different testing modalities (e.g., immunohistochemistry assays to measure BRAF protein expression) that are not represented here.

The ensemble dimensions and modalities represented in this landscape thus reflect one set of assumptions about how to summarize the evidence and provide a compelling proof of concept. Whereas this set of assumptions limits our interpretation, the evidence landscaping approach does not depend on those assumptions. Future iterations of an evidence landscaping tool could, for example, allow the user to modify and cluster the axes differently or change the color-coding scheme. This flexibility in how the ensemble space is constructed is arguably an advantage of the approach allowing the scientific community to explore the effects that different assumptions may have on the apparent contours of the evidence landscape.

Achieving the goals of precision medicine remains a challenge. The interventions in this domain are complex ensembles of technologies, procedures, and the clinical characteristics of patients. Efficiently resolving uncertainty about each of the ensemble components will require systematic experimentation and coordination. We have presented an approach that can help in this task—a principled method to map the accumulation of evidence in support of an ensemble. Using an evidence landscaping method to examine the case of BRAF mutation testing in metastatic melanoma, we have found some clinically important properties of the research and regulatory enterprises even for an approved targeted therapy that is widely used for treating melanoma. There may be much that remains unknown about the best way to test patients for BRAF mutations. However, we believe that folding this evidence landscaping approach into funding decisions, experimental design, evidence assessment, and regulatory decision-making would be a step forward for the field. It could help to better coordinate investigations, sharpen communications, and render regulatory decisions more transparent.

SUPPLEMENTARY MATERIALS

stm.sciencemag.org/cgi/content/full/12/540/eaaw7745/DC1
Literature search
Data extraction
Mapping the evidence landscape
Reference (*20*)

REFERENCES AND NOTES

1. S. P. Hey, A. S. Kesselheim, Countering imprecision in precision medicine. *Science* **353**, 448–449 (2016).
2. M. J. Khoury, C. M. McBride, S. D. Schully, J. P. Ioannidis, W. G. Feero, A. C. J. W. Janssens, M. Gwinn, D. G. Simons-Morton, J. M. Bernhardt, M. Cargill, S. J. Chanock, G. M. Church, R. J. Coates, F. S. Collins, R. T. Croyle, B. R. Davis, G. J. Downing, A. D. Ross, S. Friedman, M. H. Gail, G. S. Ginsburg, R. C. Green, M. H. Greene, P. Greenland, J. R. Gulcher, A. Hsu, K. L. Hudson, S. L. R. Kardia, P. L. Kimmel, M. S. Lauer, A. M. Miller, K. Offit, D. F. Ransohoff, J. S. Roberts, R. S. Rasooly, K. Stefansson, S. F. Terry, S. M. Teutsch, A. Trepanier, K. L. Wanke, J. S. Witte, J. Xu, The scientific foundation for personal genomics: Recommendations from a National Institutes of Health–Centers for Disease Control and Prevention multidisciplinary workshop. *Genet. Med.* **11**, 559–567 (2009).
3. D. F. Hayes, J. Allen, C. Compton, G. Gustavsen, D. G. B. Leonard, R. McCormack, L. Newcomer, K. Pothier, D. Ransohof, R. L. Schilsky, E. Sigal, S. E. Taube, S. R. Tunis, Breaking a vicious cycle. *Sci. Transl. Med.* **5**, 196cm6 (2013).
4. V. Prasad, T. Fojo, M. Brada, Precision oncology: Origins, optimism, and potential. *Lancet Oncol.* **17**, e81–e86 (2016).
5. R. M. Simon, S. Paik, D. F. Hayes, Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Natl. Cancer. Inst.* **101**, 1446–1452 (2009).
6. A. Vivot, I. Boutron, G. Béraud-Chaulet, J.-D. Zeitoun, P. Ravaud, R. Porcher, Evidence for treatment-by-biomarker interaction for FDA-approved oncology drugs with required pharmacogenomic biomarker testing. *Sci. Rep.* **7**, 6882 (2017).
7. E. D. Saad, X. Paoletti, T. Burzykowski, M. Buyse, Precision medicine needs randomized clinical trials. *Nat. Rev. Clin. Oncol.* **14**, 317–323 (2017).
8. L. M. McShane, D. G. Altman, W. Sauerbrei, S. E. Taube, M. Gion, G. M. Clark; Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics, Reporting recommendations for tumor marker prognostic studies (REMARK). *J. Natl. Cancer Inst.* **97**, 1180–1184 (2005).
9. F. Andre, L. M. McShane, S. Michiels, D. F. Ransohoff, D. G. Altman, J. S. Reis-Filho, D. F. Hayes, L. Pusztai, Biomarker studies: A call for a comprehensive biomarker study registry. *Nat. Rev. Clin. Oncol.* **8**, 171–176 (2011).
10. U.S. Food and Drug Administration Website; https://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm301431.htm [retrieved 3 January 2018].
11. P. B. Chapman, A. Hauschild, C. Robert, J. B. Haanen, P. Ascierto, J. Larkin, R. Dummer, C. Garbe, A. Testori, M. Maio, D. Hogg, P. Lorigan, C. Lebbe, T. Jouary, D. Schadendorf, A. Ribas, S. J. O'Day, J. A. Sosman, J. M. Kirkwood, A. M. M. Eggermont, B. Dreno, K. Nolop, J. Li, B. Nelson, J. Hou, R. J. Lee, K. T. Flaherty, G. A. McArthur; BRIM-Study Group, Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* **364**, 2507–2516 (2011).
12. A. Hauschild, J.-J. Grob, L. V. Demidov, T. Jouary, R. Gutzmer, M. Millward, P. Rutkowski, C. U. Blank, W. H. Miller Jr., E. Kaempgen, S. Martín-Algarra, B. Karaszewska, C. Mauch, V. Chiarion-Sileni, A.-M. Martin, S. Swann, P. Haney, B. Mirakhur, M. E. Guckert, V. Goodman, P. B. Chapman, Dabrafenib in BRAF-mutated metastatic melanoma: A multicentre, open-label, phase 3 randomised controlled trial. *Lancet* **380**, 358–365 (2012).
13. M. Colombino, M. Capone, A. Lissia, A. Cossu, C. Rubino, V. De Giorgi, D. Massi, E. Fonsatti, S. Staibano, O. Nappi, E. Pagani, M. Casula, A. Manca, M. Sini, R. Franco, G. Botti, C. Caracò, N. Mozzillo, P. A. Ascierto, G. Palmieri, BRAF/NRAS mutation frequencies among primary tumors and metastases in patients with melanoma. *J. Clin. Oncol.* **30**, 2522–2529 (2012).
14. K. T. Flaherty, I. Puzanov, K. B. Kim, A. Ribas, G. A. McArthur, J. A. Sosman, P. J. O'Dwyer, R. J. Lee, J. F. Grippo, K. Nolop, P. B. Chapman, Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med.* **363**, 809–819 (2010).
15. M. Saint-Jean, G. Quéreux, J.-M. Nguyen, L. Peuvrel, A. Brocard, A. Vallée, A.-C. Knol, A. Khammari, M. G. Denis, B. Dréno, Is a single BRAF wild-type test sufficient to exclude melanoma patients from vemurafenib therapy? *J. Invest. Dermatol.* **134**, 1468–1470 (2014).
16. H. Ledford, Cutting-edge cancer drug hobbled by diagnostic test confusion. *Nature* **556**, 161–163 (2018).
17. F. Spagnolo, P. Ghiorzo, L. Orgiano, L. Pastorino, V. Picasso, E. Tornari, V. Ottaviano, P. Queirolo, BRAF-mutant melanoma: Treatment approaches, resistance mechanisms, and diagnostic strategies. *Onco. Targets Ther.* **8**, 157–168 (2015).
18. S. Anderson, K. J. Bloom, D. U. Vallera, J. Rueschoff, C. Meldrum, R. Schilling, B. Kovach, J.-R. Lee, P. Ochoa, R. Langland, H. Halait, H. J. Lawrence, M. C. Dugan, Multisite analytic performance studies of a real-time polymerase chain reaction assay for the detection of BRAF V600E mutations in formalin-fixed, paraffin-embedded tissue specimens of malignant melanoma. *Arch. Pathol. Lab. Med.* **136**, 1385–1391 (2012).
19. S. P. Hey, C. Weijer, M. Taljaard, A. S. Kesselheim, Research ethics for emerging trial designs: Does equipoise need to adapt? *BMJ* **360**, k226 (2018).
20. S. P. Hey, C. M. Heilig, C. Weijer, Accumulating evidence and research organization (AERO) model: A new tool for representing, analyzing, and planning a translational research program. *Trials* **14**, 159 (2013).

Funding: S.P.H., V.P., and A.S.K. have received funding from Arnold Ventures, a philanthropic organization that funds research in criminal justice, education, health, and public finance. **Competing interests:** V.P. has received book royalties and is a writer for Medscape.

10.1126/scitranslmed.aaw7745

Citation: S. P. Hey, C. V. Gerlach, G. Dunlap, V. Prasad, A. S. Kesselheim, The evidence landscape in precision medicine. *Sci. Transl. Med.* **12**, eaaw7745 (2020).

Science Translational Medicine

The evidence landscape in precision medicine

Spencer Phillips Hey, Cory V. Gerlach, Garrett Dunlap, Vinay Prasad and Aaron S. Kesselheim

Sci Transl Med **12**, eaaw7745.
DOI: 10.1126/scitranslmed.aaw7745