Published in final edited form as:

Ann Intern Med. 2017 October 03; 167(7): 511–512. doi:10.7326/M17-1733.

Population Health Science and the Challenges of Prediction

Sandro Galea, MD, DrPH and

Boston University, Boston, Massachusetts

Katherine M. Keyes, PhD

Columbia University, New York, New York

Peering into the future to accurately predict health outcomes is challenging and imprecise. Yet, this has not stopped the development of a broad array of models to predict future disease. The widespread availability of regression modeling approaches has made it seem easy to develop prediction models. A substantial body of literature documents the development, calibration, and validation of risk prediction models using large data sets. Typically, these models use individual patient risk factors to estimate the probability of a future health outcome. Patient-friendly Web sites provide tools that enable patients to calculate their own risk for such conditions as cardiovascular disease or cancer, in theory to motivate them to improve their risk profiles. However, the prediction modeling enterprise faces substantial challenges that seldom attract as much attention as they should.

Prediction models rest on the notion that we have enough information to develop a statistical picture of a patient's future health status. Taken at face value, this proposition is appealing because it builds on our understanding of a broad set of risk behaviors and an emerging set of genetic and other biological markers known to be associated with disease at the population level. However, 4 core factors limit the accuracy of predictive modeling: the strength of the relationship between the risk factors and the outcome, the scale with which we assess "strength" of association, the base prevalence of the variables in question, and the co-occurring factors that interact with the risk factor sources (1). Dalton and colleagues' report in *Annals* (2) provides a sobering reminder of the limitations of predictive models that fail to consider base prevalence and co-occurring factors. These researchers show that a neighborhood disadvantage index (NDI) outperformed the Pooled Cohort Equations Risk Model (PCERM) of the American College of Cardiology and American Heart Association by a factor of 3.

This work demonstrates the limitations of the PCERM and similar models that do not include data about patients' ecologic circumstances. The NDI likely captures a breadth of factors that are not readily quantifiable in data sets typically used to validate risk models. For example, the authors note that neighborhood walkability and the availability of healthy

Requests for Single Reprints: Sandro Galea, MD, DrPH, Boston University of Public Health, 715 Albany Street, Boston, MA 02118; sgalea@bu.edu.

Current Author Addresses: Dr. Galea: Boston University School of Public Health, 715 Albany Street, Boston, MA 02118.
Dr. Keyes: Columbia University Mailman School of Health, 722 West 168th Street, 724, New York, NY 10032.

Disclosures: Authors have disclosed no conflicts of interest. Forms can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M17-1733.

food might well be reflected in a general neighborhood disadvantage measure. Even more important, the dramatically better predictive performance of the NDI compared with the PCERM highlights the challenges facing any effort toward individual disease prediction.

Two conceptual frameworks in the population health literature capture the essential problem with which any predictive model must grapple. The life course framework proposes that health is produced through a complex set of factors dynamically interacting across the life span and influenced by intergenerational forces (3). Operationally, this means that an individual's health is forged by factors that antecede his or her conception, are shaped during the intrapartum period, are strongly influenced by childhood factors, and then are modified throughout adulthood. The life course model, now well-established in the population health literature, suggests that the range of factors that predict one's health vastly transcend the point-in-time snapshots offered by individual behaviors in the present or biomarkers that may represent health and well-being at any particular time point (3). It also amply illustrates that the success of any prediction model depends on inclusion of the many factors that influence the production of health, which may reach back into the past and may change continually as the individual interacts with his or her environment. This highlights the challenge that befalls attempts to predict health outcomes from a neat, small set of variables.

The framework of multilevel modes of disease causation posits that the production of health rests not only with the individual (4). Although ultimately all health is "under the skin," mediated by individual behavior and biology, both behavior and biology are shaped by a person's social networks, neighborhoods, policies that influence access to healthy resources, and social norms that guide behaviors. Dalton and colleagues (2) illustrate this concept nicely through the predictive power of the NDI. Clearly, neighborhood disadvantage produces a set of conditions that increase the risk faced by those living in marginalized neighborhoods in ways that an individual risk score simply cannot capture.

Further, disease often—perhaps nearly always—develops as a result of the co-occurrence of several causes. From an epidemiologic perspective, this means that each cause is insufficient by itself; several causes are necessary for disease to develop (5). The implication of this observation is that unless we understand all the factors that cause a particular disease, our understanding of the role of any one factor inevitably will be limited. In fact, the risk difference, or risk ratio, for any given factor's effect on disease is entirely a function of the prevalence of factors that interact with it (6). For example, although genes indeed may be classic "risk factors" for complex diseases, their effect matters only insofar as other factors that influence their transcription are present. Therefore, our understanding of the role of genes in disease prediction may be only as good as our understanding of these other co-occurring factors (7). The core challenge here is that we frequently do not know what these co-occurring factors are, which hampers accurate prediction.

These challenges warrant substantial humility in the development and use of predictive models and a greater focus on social, economic, and geographic assessments in these models. Nevertheless, predictive models, even if imperfect, may have some clinical utility. However, predictive models must be applied with caution. Consider the following hypothetical example: Suppose we estimate that persons with gene *X* have 5 times the risk

for disease than those who have gene *Y*. If we give the patient with gene *X* this information without any context, she might think she is doomed to develop the disease. However, if the prevalence of the condition in the population is 0.02%, her risk on an additive scale is increased only by 0.08 percentage point, perhaps an increase hardly worth worrying about. Further, if we consider the factors that co-occur with gene *X*, her risk may in fact not be increased at all.

Even if clinicians recognize the conceptual challenges behind prediction models, they may have difficulty factoring in those limitations when considering the neat percentage risk for a particular disease that a predictive model offers. Of course, if that risk is misestimated, patient management may be suboptimal. The field of predictive models in health is overdue for a conceptual and methodological rethink, building on the principles of population health science mentioned earlier. The current study by Dalton and colleagues (2) is a good first step in this direction.

References

1. Keyes KM, Galea S. Population Health Science. Oxford, UK: Oxford Univ Pr; 2016.
2. Dalton JE, Perzynski AT, Zidar DA, Rothberg MB, Coulton CJ, Milinovich AT, et al. Accuracy of cardiovascular risk prediction varies by neighborhood socioeconomic position. A retrospective cohort study. *Ann Intern Med*. 2017; 167:456–64. DOI: 10.7326/M16-2543 [PubMed: 28847012]
3. Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. Life course epidemiology. *J Epidemiol Community Health*. 2003; 57:778–83. [PubMed: 14573579]
4. Kaplan GA. What is the role of the social environment in understanding inequalities in health? *Ann N Y Acad Sci*. 1999; 896:116–9. [PubMed: 10681892]
5. Rothman KJ. Synergy and antagonism in cause-effect relationships. *Am J Epidemiol*. 1974; 99:385–8. [PubMed: 4841816]
6. Keyes KM, Galea S. Setting the Agenda for a New Discipline: Population Health Science [Editorial]. *Am J Public Health*. 2016; 106:633–4. DOI: 10.2105/AJPH.2016.303101 [PubMed: 26959265]
7. Keyes KM, Smith GD, Koenen KC, Galea S. The mathematical limits of genetic prediction for complex chronic disease. *J Epidemiol Community Health*. 2015; 69:574–9. DOI: 10.1136/jech-2014-204983 [PubMed: 25648993]