Published in final edited form as:

Personal Disord. 2019 January ; 10(1): 46–58. doi:10.1037/per0000300.

Health Risk Prediction Models Incorporating Personality Data: Motivation, Challenges, and Illustration

Benjamin P. Chapman[1,2], **Feng Lin**[1,3,4], **Shumita Roy**[5], **Ralph H.B. Benedict**[5], and **Jeffrey M. Lyness**[1,6]

[1]Department of Psychiatry, University of Rochester Medical Center

[2]Department of Public Health Sciences, University of Rochester Medical Center

[3]School of Nursing, University of Rochester Medical Center

[4]Department of Brain and Cognitive Sciences, University of Rochester

[5]Department of Neurology, University at Buffalo Medical Center

[6]Department of Neurology, University of Rochester Medical Center

Abstract

The age of "big data" in health has ushered in an era of prediction models promising to forecast individual health events. While many models focus on enhancing the predictive power of medical risk factors with genomic data, a recent proposal is to augment traditional health predictors with psychosocial data, such as personality measures. In this paper we provide a general overview of the medical risk prediction models, then discuss the rationale for integrating personality data. We suggest three principles that should guide work in this area, if personality data is ultimately to be useful within risk prediction as it is actually practiced in the health care system. These include a) prediction of specific, priority health outcomes; b) sufficient incremental validity beyond established biomedical risk factors; and c) technically responsible model-building that does not overfit the data. We then illustrate the application of these principles in the development of a personality-augmented prediction model for the occurrence of Mild Cognitive Impairment (MCI), designed for a primary care setting. We evaluate the results, drawing conclusions for the direction an iterative, programmatic approach would need to take to eventually achieve clinical utility. While there is great potential for personality measurement to play a key role in the coming era of risk-prediction models, the final section reviews the many challenges that must be faced in real-world implementation.

Keywords

personality; personalized medicine; medical prediction models; mild cognitive impairment

Corresponding Author: Ben Chapman, PhD MPH, Department of Psychiatry, University of Rochester Medical Center, 300 Crittenden Blvd., Rochester, NY 14642, 585-319-9019, ben_chapman@urmc.rochester.edu.

Conflicts of Interest:
None to declare

Introduction

Based on at least three decades of work demonstrating relationships between personality and health outcomes, interest is now coalescing around how personality data could be practically used to improve people’s health. When a set of risk factors is discovered, the most common instinct is to seek to modify them. This thinking underlies recent considerations of how health-relevant personality traits might be targeted via intervention, as other papers in this special issue describe. While strong arguments exist for that prospect, there are also some scenarios in which trait change may be less tractable (Chapman, Hampson, & Clarkin, 2014; Mroczek, 2014) and literature on the predictive power of personality for health outcomes predates that calling for interventions. Some traits may be less amenable to change, some persons unable or unwilling to change, available time frames too short for meaningful trait change, and external (lack of a supportive environment) or internal (neurodegenerative disease) circumstances may preclude the desired changes. In yet other cases, it may be possible to change a trait, but such change may not alter the likelihood of the focal health outcome. These cases take one of two forms: first, the health damaging effects of a particular trait have already accumulated, particularly if it has been operating over the better part of a person’s lifespan. A sensation-seeking extravert who takes up smoking at a young age and continues it for decades may see a reduction in these traits, naturalistic or otherwise, in middle or later life. Neither the cessation of smoking nor the remediation of years of accumulated lung damage may result. Second, a personality trait may not be causally related to an outcome, but instead may proxy another etiologic factor, such as a genetic disposition or environment.

In light of these considerations, should we abandon hope of utilizing personality information to improve health outcomes? We think not, and in this paper describe an alternative way in which personality can be practically used to improve health outcomes. The alternative to direct intervention hinges on leveraging the long-term predictive power of personality measures for health outcomes (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007) in tools of potential use within health care clinics. These tools, variously called risk prediction models, risk scores, risk calculators, prognostic models, and prediction models, are empirically based forecasts of the likelihood a given patient will experience a particular health outcome over some defined time horizon. The first part of this paper provides an overview of health prediction models, including their place within the recent precision medicine movement. The rationale for augmenting these models with personality measures is laid out, and three principles are proposed to determine whether or how personality data might be usefully integrated into such models. The second section of the papers features a concrete illustration of how a standard health risk model based on demographic and medical data could be augmented by personality measures. The third and final portion of the paper focuses on the challenges inherent in the idea, or indeed in any effort to integrate personality into actual health care practice.

Health Risk Prediction

There are several advantages to quantitative prediction tools that accurately foretell the occurrence of a disease, its prognosis or course, or an individual’s likelihood to respond to a certain treatment. Such tools a) enable patients and their families to make more informed decisions about treatment and prevention (for instance, balancing the side-effects of a prevention regimen against the individual's likelihood of experiencing that outcome); b) help clinicians precisely tailor care by planning treatment and prevention; and c) aid health care systems in allocating resources to patients most at risk for an outcome. The latter process is known in medicine as risk stratification, or the ordering large numbers of patients in strata reflecting increasing levels of risk for the health outcome.

Prediction models in clinical medicine are not new. For instance, one of the most widely used prediction models is the Framingham Risk Score (FRS). The FRS takes data on cardiovascular factors such as smoking or obesity, and based on a validated logistic regression model a cardiovascular outcome such as stroke or myocardial infarction, produces a probability of that outcome (available at http://www.cvriskcalculator.com/). This probability then informs treatment. For instance, The American College of Cardiology and American Heart Association recommend that statin treatment be initiated if a risk score of > 7.5% chance of stroke or myocardial infarction in the next ten years is achieved for 40–75 year patients free from cardiovascular disease (Goff et al., 2013).

However, predictive medicine has entered a new age of potential due to the advent of "Big Data," which involves a tidal wave of novel biomarkers and the burgeoning architecture of information management systems such as electronic medical records (EMRs). Classic approaches to data-driven prediction have appeared under the headings of "prognostic models" or "prognostic medicine"[1] (Royston, Moons, Altman, & Vergouwe, 2009). Those models typically employed information readily available to health professionals in their clinics, such as demographics and medical factors assessed in the office. The goal of any prediction model is to combine available data on risk factors into a single index, usually called a risk score, which conveys some information about the likelihood of experiencing an outcome (see Royston et al., 2009). To achieve sufficient predictive validity for use in actual health care clinics, successful risk models typically go through many iterations of validation and revision in independent samples reflecting their target population.

The concept of a risk-score from a medical prediction model is not unlike that of a score from a multi-item psychological scale: several relevant measurements (for instance the presence of diseases or health risk behaviors) are added up, possibly using weights. The weights must be based on some reliable empirical analysis of relevant data if the risk score is to be at all useful. Health care decision support software, either free-standing or integrated directly into EMRs, can now computes risk scores automatically if the relevant input data has been collected at the point of clinical care. There are many databases of "risk calculators" for the practicing physician to use. For example, one popular web site is MDcalc (www.mdcalc.com), a subscription-based web warehouse of predictive models (and other clinical tools). Health care providers can search for risk models relevant to any given patient, input the relevant information, and receive an immediate score (as well as other

[1] The term prognostic literally suggests predicting the course of a disease once it has been diagnosed. However, the intent is to predict any health outcome, e.g. a incidence of a disease that has not yet been diagnosed, so the term tends to be used synonymously with "predictive".

information on the particular prediction model, including its supporting studies). This has an indirect benefit of permitting more sophisticated statistical prediction models, since there is no need to over-simplify the resulting formula of the risk score for hand calculation[2].

The advent of genomics produced a more widely recognized application of predictive models known as “personalized medicine” (Hamburg & Collins, 2010). Much of the early work in personalized medicine involved identifying patient genetic markers predicting response to particular forms of cancer treatment. The term expanded quickly to encompass individualized approaches to patient care based on genomic assessment of disease risk. Predictive models of disease incidence in this arena have generally been based on culling through a large number of single nucleotide polymorphisms (SNPs) to identify a small, highly predictive subset which are then combined into a single estimate of genetic risk, called a polygenic risk score (e.g., Peterson et al. 2011). This process of constructing such a risk score involves quantitative methods suited to large numbers of predictors, known as machine learning (more aptly called statistical learning; see Hastie, Tibshirani, & Friedman, 2009). The branch of machine learning used for prediction models can be thought of conceptually as a regression-modeling framework tailored to deal with a large or complicated predictor set.

Modifiability and Causality

Polygenic risk scores illustrate a key point of prediction models—that risk factors need not be modifiable to be incorporated into patient care, prevention, and treatment planning. That point is particularly relevant for personality traits, the plasticity of which may vary across different persons, periods of life, and traits. The reason is that modifiability is irrelevant in the question of whether some measurement in the present, be it genotype or phenotype, predicts something in the future. The causal status of a risk factor is also irrelevant from this prediction perspective, since the risk factor is simply a measurement that conveys information about the likelihood of a future outcome—the proverbial “canary in the coal mine”. The measurement may be an easily attainable proxy for some other truly causal factor, an amalgam of causal and non-causal factors, or an indirect or distal factor on a complex causal chain. For instance, a particular personality trait score may be the result of one or more genetic and/or environmental factors that are also causes of a disease outcome. In this case, the trait score does not cause the outcome, but is an easily measurable epiphenomenon of an underlying cause of the outcome and therefore provides predictive leverage. Of course, if one is to move a step beyond the “early warning” goal of risk prediction and wishes to use a component of a risk score as an intervention target, then that component ought to be causally linked to the outcome.

From a clinical perspective, it is paramount to first identify who is at risk for an outcome. Without accurate predictions, it is not even possible to know in whom to undertake an intervention with established effectiveness or target pathways known to be causal. Of course, these are not “either/or” issues: modifiable causal predictors provide immediate hints to risk

[2] For instance, many models simply rounded the regression weights needed for specific components of the risk score to simple integer values, to make hand calculation by busy clinic staff easier.

reduction, and the causality and modifiability of many predictive factors is often not entirely known.

Precision Medicine and Personality

The expansion of prediction models to include many levels and types of information underlies the new precision medicine movement of the US National Institutes of Health (NIH). Precision medicine is defined as “an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person" (NIH, 2016). A decade earlier, personality was pinpointed as an “intermediate phenotype” at the nexus of these three domains (Institute of Medicine, 2006). These considerations have given rise to suggestions that personality traits be studied as adjunctive markers in health prediction models, as a phenotypic complement to genotypic data (Chapman, Roberts, & Duberstein, 2011; Chapman, Hampson, & Clarkin, 2014).

That notion is predicated on three findings in the literature. First, the predictive power of personality for consequential life outcomes (including health) has been repeatedly demonstrated, and is on par with other factors such as IQ or socioeconomic status (SES; Roberts, et al. 2007). It stands to reason that a construct touted for its predictive power would be a prime candidate for consideration in medical risk prediction models. Second, the personality traits and facets encompassed by the Big Five framework tap propensities toward health behaviors, physiological processes relevant toward health, and indeterminate genetic factors that may have health consequences. Some of this information will encompass specific risk factors—for instance, Conscientiousness is linked to better diet and lower likelihood of smoking, which lead to lower blood pressure (Bogg & Roberts, 2004)—but some of the information contained in personality measures will provide unique, incremental predictive value beyond specific risk factors. If an individual’s risk score is high because of a particular profile of elevations (or decrements) in key personality facets, then this information also may provide some insight on tailoring prevention or intervention (for instance, anxiolytics when trait anxiety drives up risk for an outcome). Third, studies prospectively predicting health outcomes are nearly all based on patient self-report scales of personality[3]. Personality items with predictive value can thus easily be incorporated in patient health history questionnaires (paper and pencil or tablet-based), online patient portals, or in similarly expedient, non-invasive, and inexpensive means. Formal testing by licensed psychologists (cost and time prohibitive in most medical settings) is not required.

Incorporating Personality Information in Medical Prediction Models

Exactly which health outcomes might be better predicted with the assistance of personality data is largely an open question. The use of personality data in a prediction model is best justified when there is at least some literature documenting basic associations in traditional research designs (i.e., studies that do not specifically develop or validate a prediction model). In a recent proof of concept study, we identified a small set of specific personality scales and SES markers that enhanced the performance of the Charlson Comorbidity Index,

[3] This is not to discount the importance of informant report or other forms of personality measurement—indeed, such information may be extremely predictive—but studies of health outcomes incorporate these forms of measurement only rarely.

a classic risk model for all-cause mortality based on demographics and chronic disease (Chapman, Weiss, & Duberstein, 2015). That study focused on facet level scales rather than broad, composite scores on domains such as Neuroticism and Extraversion, as narrower trait measurements offer better predictions under some conditions. Another approach has been to construct predictive scales from the item level itself (Chapman, Weiss, & Duberstein, 2016; Weiss, Gale, Batty, & Deary, 2013). Working directly at the item level treats personality inventories as a so-called "SNP chip" or measurement of a large number of SNPs. Describing this approach, Weiss and colleagues (2013) coined the term "Questionnaire-wide Association Study" (QWAS) as an analogy to the Genome Wide Association Study (GWAS). The item- or QWAS level of analysis trades the advantage of conceptual trait constructs, represented by narrow trait scales, for the greatest level of granularity possible in personality measurement. Depending on the context, either may be defensible. In our example illustration below, we identify a middle ground between pure item-analysis and pure facet-scale analysis, based on items and item parcels that have known mappings to distinct trait constructs. We now review three key principles for personality-informed prediction models.

Three Principles For Personality Informed Prediction Models

1) Clinically Relevant Endpoints—First, the prediction model itself must provide risk estimates for an endpoint of clinical relevance. Most of clinical endpoints of interest have at least some recognized prevention and early intervention strategies, and are thus the focus of health care providers in clinics. Because health conditions require formal definitions, they are often categorical in nature. For instance, systolic and diastolic blood pressures are continuous measures, but hypertension is defined by a threshold ratio of the two (140/90). Thus, models could be constructed producing a predicted amount of increase in blood pressure (a continuous outcome), but that increase might or might not exceed the defined threshold for hypertension. This issue is not so much over whether it is statistically better to use continuous or categorical measures, as a matter of working within the discrete endpoints around which healthcare decision-making, research, and treatment is centered.

2) Added Value Over Biomedical Information—Efforts to integrate personality information into medical prediction models must similarly work within the confines of established health risk factors, and existing models based on those risk factors. No medical practice would use a risk score for Myocardial Infarction, for instance, based solely on personality items when lipids, body mass index, and other relevant biomedical data is routinely available. Thus, the goal is to identify personality measurements that enhance prediction beyond the components found in existing risk models. If a personality measure provides no added predictive power beyond information already available, it is unlikely to be of any practical use.

Improvement of a model can come in different forms (Steyerberg et al., 2010). Calibration refers to whether the model systematically over or under-predicts the outcome in general, or within certain regions of the risk spectrum. For instance, the existing model may be highly accurate at extremely low and high risk, but poor between. Discrimination refers to how well the model distinguishes "cases", or those who experience the outcome, from non-cases or "controls" who do not. It is based on the separation of the risk scores across these two groups, and may be good even if calibration is bad. For instance, cases may always have a higher risk score, but miscalibration may lead to systematic overestimation of everyone's risk. Overall measures of model performance combine information on calibration and discrimination.

In practice, risk scores will be categorized to abet decision making about whether to refer patients for further testing, initiative interventions, and so forth. Thus, it is also useful to consider whether added data improves false positive or false negative rates at cut-points likely to be used in clinical practice for a given screening strategy. One common screening strategy is to set a very low cut-point, to catch as many of those who actually experience the outcome, or cases, as possible (Gordis, 2005). Setting a low cut-point to maximize sensitivity (the proportion of true cases identified) but will necessarily bring with it false positives, and one would want added information to reduce the number of false positives at a given sensitivity. The opposite strategy would be to set a very high cut-point. This would capture nearly everyone who doesn't eventually experience the outcome, or non-cases (Gordis, 2005). Of course, maximizing specificity (true non-cases identified as non-cases) would also miss those who truly experience the outcome, or false negatives. If a risk prediction program prioritizes specificity, one would want added information such as personality data to reduce false negatives at a given specificity level. The cost of a single missed case (false negative) vs. false positive, as well as the total number of each (a function of the base rate of the outcome) will influence whether a health care system prefers to maximize sensitivity or specificity in a given risk-screening initiative. Any new marker considered in the risk model may facilitate one, both, or neither of these practical goals.

3) Technical Soundness—Just as the evaluation of whether or not personality measurements improve an existing prediction model is complex, so is the construction of the model itself. This is a statistical task, and there are many places it can go wrong. A predictive model must produce risk scores that have predictive value beyond the sample in which the model was developed. This is a universal imperative for prediction models, but one that is sometimes neglected. The technical term for a model that performs poorly out-of-sample, and is thus not generalizable or useful, is "overfitting." Overfitting occurs when too many predictors are included, or parameter estimates that are too larger, or parameters are too complex (i.e., many non-linear terms and interactions). Such a model looks like it predicts the outcome well in the data at hand, but that performance is partially driven by idiosyncrasies of that sample. The same model may predict the outcome quite poorly in a different sample lacking those idiosyncrasies. Many machine-learning algorithms are designed specifically to prevent this. Conceptually, these algorithms work by "underfitting" the model in the development sample, using internal cross-validation (see Hastie et al., 2009 for technical exposition, Chapman et al., 2016 for an introduction for psychologists). Of course, one does not want to underfit the data at hand too much, or this too could result in poor prediction. Ideally, a completely independent external validation sample is used for a final assessment of model performance. This is rare in practice, and in fact even the use of internal cross-validation or other low-cost strategies to prevent overfitting is rather scant (see, e.g., Royston et al., 2013).

Cross-validation and other techniques in machine learning are generally not based on p-values. This does not mean that they fail to distinguish between useful predictors and non-useful predictors; quite the contrary. Machine learning algorithms are simply designed to build the best generalizable prediction from the predictor set, not perform null-hypothesis tests on individual predictors. One might think of the regression weights they produce as measures of "effect size." Typically, if a variable is not useful, it will be discarded entirely or receive a very small weight. Certain methods, such regularization (utilized in the example below), shrink coefficients to a degree roughly proportional to their standard error. Machine learning methods are generally quite different from traditional methods such as mediation models, but can be thought of as efforts to estimate the total association between a predictor and outcome, whatever pathways or mechanisms might be responsible for it. Classical methods such as logistic regression have also been employed in model construction for years (Royston et al., 2009). Their advantages are familiarity and some idealistic statistical properties of maximum likelihood estimates (MLE). Ultimately the choice of approach for any given application depends on a variety of factors, but the end goal is a model with generalizable predictive validity.

In the next section, we illustrate the application of these three criteria—clinically meaningful outcomes, incremental prediction accuracy, and strategies to avoid overfitting--in an example. We then conclude with a general discussion of the promises and pitfalls inherent in efforts to translate personality research in clinically useful healthcare tools.

Predicting Mild Cognitive Impairment

Dementia and Early Risk Prediction

Dementia is a general term that refers to a variety of cognitive disorders. Alzheimer's Disease and related dementias are officially listed as the sixth leading cause of death in the US (Centers for Disease Control [CDC], 2015), but may in fact be as high as the third or fourth: mortality burden is underestimated because they are often not recorded or recorded only as contributing causes on death certificates (James, Leurgans, Herbert, Scherr, Yaffe, & Bennett, 2014). Economic burden is also extraordinarily high, and dementia incidence is projected to increase in coming decades (CDC, 2015). Although the development of treatments for AD and related dementias is a major priority, no therapies exist which can alter the gradual progression of the disease. Therefore it has become imperative to predict signs of dementias as early as possible, to provide ample lead-time for preparation, planning, and efforts to slow progression.

Mild Cognitive Impairment (MCI) is a condition for which different sets of diagnostic criteria exist (Albert et al., 2011). Common features across all classification schemes include a) the maintenance of everyday activities of daily living, combined with b) a change for the worse in one or more domains of cognitive functioning, such as memory or executive function. Another common feature of the condition is self- reported cognitive decline, although there is no consensus that this should be a required diagnostic feature. Because it reflects less severe impairment, MCI is not the same as dementia, but can progress to dementia. Although different subtypes of MCI are not always distinguished in practice, amnestic MCI (aMCI) is a common variety of MCI involving memory impairment. The term "non-amnestic MCI" is used to refer to cases in which cognitive impairment in other (non-memory) domains is evident. For instance, "executive MCI" (eMCI) involves deficits in executive function (Reinvang, Grambaite, & Espeseth, 2012). Different varieties of MCI may presage different forms of dementia. Ultimately, the transition from normal cognitive function to possible MCI a very early prediction target. An ideal risk-prediction paradigm would identify persons who are as yet without cognitive deficits, but who evidence a higher than average probability of developing some type of MCI in the future.

Few formal risk models exist for MCI or for its subtypes. Most risk prediction models focus on dementia itself (Stephan et al., 2010), with reported models for MCI based on age, education, and health factors—in particular, vascular risk factors such as hypertension, and hypercholesterolemia (Pankratz et al., 2015; Unverzagt et al., 2011). A second kind of MCI risk model focuses on Magnetic Resonance Imaging (MRI), cerebrospinal fluid (CSF), and other more specialized biomarker measurements such as Apolipoprotein E (Heister et al., 2011; Kantarci et al., 2013). Biomarkers offer a promising means of predicting the onset of cognitive deficits at preclinical phases of diseases (Sperling & Johnson, 2013), but typically involve referral to specialty clinics and/or more expensive or invasive procedures.

Current possibilities for MCI risk prediction consist of a) "first-line" risk models that could be computed in general medical settings based on standard health data and demographics; and b) models based on more expensive and nuanced information such as brain imaging. A comprehensive early detection initiative would ideally leverage both layers. For instance, a serial prediction program might maximize sensitivity at an initial screening stage with widely available data in general practices, capturing the largest number of potential cases. Those deemed at risk in this stage could then be referred to specialty clinics for more expensive and invasive biomarker data, which could then be used to rule-out false-positives from the initial stage.

The first criterion for personality-informed risk prediction models—an endpoint of major public health importance—is therefore satisfied in MCI. In this example application, the cognitive battery is more sensitive to possible eMCI, though cannot fully distinguish between subtypes. Were the models to be iteratively refined, the next study would probably attempt to define the endpoint with a more extensive battery. Criterion two, that personality adds some predictive utility beyond established risk factors, presupposes as a start that personality prospectively predicts cognitive outcomes. Such a base of evidence exists, with meta-analyses linking personality traits to both formal dementia (Low, Harrison, & Lackersteen, 2013) and cognitive decline (Luchetti, Terracciano, Stephan, & Sutin, 2016). Existing prediction models based on standard health (Pankratz et al., 2015; Unverzagt et al., 2011) as well as imaging data (Heister et al., 2011; Kantarci et al., 2013; see also Stephan et al., 2010) also appear to perform modestly, suggesting that there is "room for improvement" from auxiliary forms of data. Therefore, a plausible basis exists to suspect that personality, in this case, might successfully augment biomedical risk factors in outcome prediction.

Why look to non-cognitive data, rather than traditional cognitive screening instruments? The entire point of using information other than cognitive screens is to anticipate or predict the very change captured by screeners. In other words, a screener is a case-finding tool for identifying potentially existing deficits. The goal of predictive models is to forecast the appearance of these deficits, before they actually occur and are detected by screeners. This is the crux of the difference between predictive models and screening procedures. As an analogy, carriers of the BRCA gene are at heightened risk for breast cancer, and this knowledge is useful in identifying elevated risk of potential future disease. Waiting instead for a suspicious mammogram result (a screening test) detects possible cases of breast cancer that have actually occurred. In general, the tasks of predicting onset, assessing severity of an existing condition, and quantifying its likely course or progression are distinct and often require complementary sources of information. The utility of different forms of any data is tied to what one hopes to find out. For instance, the type and rate of personality change occurring after a dementia diagnosis may be a marker of progression and prognosis (i.e., whether behavioral disturbance will occur or worsen). In the present context, we focus exclusively on whether personality information in cognitively intact older individuals can predict the incidence of potentially significant cognitive change over a relatively short (2–3 year average) time span. The third criterion involves steps to reduce model overfitting, and is described below.

Participants and Procedure

Primary care clinics represent the first (and often only) point of contact for persons presenting with dementing disorders, and have been an increasingly popular site for early detection or risk-prediction efforts. Such settings also tend to have data available on a variety of general health factors and demographics potentially useful in MCI prediction. We thus utilized a large longitudinal sample of persons 65 and older presenting at primary care clinics in the Rochester, NY area. The sample and study procedures are described in detail in Chapman, Roberts, Lyness, and Duberstein (2013) and all procedures were approved by the local institutional review board. Briefly, older persons were recruited from the waiting rooms of primary care clinics, with consenting individuals interviewed at home visits by research assistants at baseline and then yearly thereafter throughout the duration of study funding, which spanned five years. Under rolling enrollment, an individual recruited in the first year could have up to four yearly follow-up assessments, while a person recruited in the fourth year would be eligible only for a one-year follow-up. Of individuals at baseline, 67% (N = 508) completed the NEO-Five Factor Inventory (NEO-FFI) of personality. A total of 385 persons with personality data scored in the unimpaired range on all three measures of cognition (see below) and represented the cognitively normal baseline cohort for which 4-year possible eMCI risk prediction models were developed. This sample was 61% female, with a mean / standard deviation (M / SD) age of 74.4 / 6.3, 14.5 / 2.3 years of education, and was roughly 96% white.

Measures

MCI—For the purposes of this demonstration, we defined an endpoint of possible MCI based on normative criteria for impairment on at least one of three cognitive measures. A one SD deficit relative to normative data is one suggested cutoff used to establish impairment in MCI (Cook et al., 2013). The first test was the Mattis Dementia Rating Scale (MDRS) Initiation-Perseveration (IP) scale (Mattis, 1988). The MDRS-IP test involves spontaneous verbal generation and accurate phoneme repetition, as well as production of prescribed graphomotor and motor patterns. The MDRS-IP primarily tests executive function, requiring also verbal, visual, and motor function. A score of 32 corresponds to one standard deviation (SD) below the Mayo Older Americans Normative Studies normative data (a score of ranging from the 11–18th percentile depending on exact age; see Lucas et al., 1998). Second, the Trail Making Tests (TMT) parts A and B (Reitan, 1958) were administered. The TMT is one of the most well-known and widely used neuropsychological tests, requiring planning and sequencing, set-shifting, and visuomotor skills. While the raw A and B scores have been commonly used, they are heavily influenced by processing speed declines seen in normal aging and more recent studies suggest that the ratio of TMT B / TMT A better reflects actual impairment (e.g., see Lamberty & Axelrod, 2006 for a review). The B/A score provides within-person standardization for visuomotor processing speed, and is also less sensitive to education (Christidi, Karaizou, Triantafyllou, Anagnostouli, Zalonis, 2015; Lamberty & Axelrod). An initial B/A score of 3 or higher was suggested to indicate impaired executive function, although other normative data suggest that this may be low (Drane, Yuspeh, Huthwaite, & Kingler, 2002). Scores of 4 or higher, however, correspond to about one SD or more above the mean in age 60+ normative samples (Drane et al., 2002, Hester, Kinsella, Ong, & McGregor, 2005), as well as the average score in several impaired samples (see Lamberty & Axelrod), and were thus used here to define impairment. Finally, the Mini-Mental status Exam (MMSE) is a well-known assessment of general cognitive function, with scores of 24 used to indicate impairment (Folstein, Folstein, & Hugh, 1975). MCI and dementia classifications are often qualified as possible, probable, etc. In this case, impairment in one or more of these measures is best regarded as possible MCI.

Demographic factors—Age, gender, and education were assessed at baseline via self-report. Demographic factors were coded so that zero represented a meaningful (and non-sample dependent) reference value, in order to interpret risk scores on the relative hazard metric of the Cox model (see below): Age was centered at 75, education at 14 years (an associate's degree), and men were coded zero and women one. Demographic factors were included in all risk scores, given their importance in cognitive outcomes and availability for use in any health setting.

Medical Risk Factors—A core set of health risk factors such as hypertension and tobacco use form the backbone of recent consensus reviews on dementia risk (Deckers et al., 2015), as well as appearing in most risk scores used in dementia prediction such as the Framingham Stroke Profile (Unverzagt et al, 2011), Framingham Vascular Risk Score and Cardiovascular Risk Factors, Aging, and Dementia Score (Kaffashian et al., 2013). In the current study, the American Heart Association's Cerebrovascular Risk Factor (CVRF) scale was completed based on medical records, including diabetes, hypertension prescription, cardiovascular disease (CVD), current and former smoking, atrial fibrillation, left ventricular hypertrophy systolic blood pressure, total cholesterol, and HDL cholesterol. The latter three are assigned ordinal categories of increasing severity (see Supplement, Table S1), with zero representing a healthy reference category; others were coded as present (one) vs. absent (zero). In addition to CVRFs, organ system ratings from the Cumulative Illness Rating Scale (CIRS; Linn, Linn, & Gurell, 1968) were also considered in developing the biomedical risk score. The CIRS quantifies an individual's overall state of health based on an analysis of organ systems, rated by a physician (JML) based on chart reviews as free from disease (0), mildly (1), moderately (2), severely (3), or extremely severely (4) burdened. CIRS ratings correlate with blind pathologist autopsy report (Conwell, Forbes, Cox, & Caine, 1993). We considered ratings from the respiratory, eyes/ears/nose/throat (EENT), upper gastro-intestinal (UGI), hepatic, endocrine, muskuloskeletal, and neurologic divisions of the CIRS. Cardiac, respiratory, and renal (closely connected to diabetes) scores were not used due to overlap with specific CVRFs.

Personality—The Big Five composite domains of Neuroticism, Extraversion, Conscientiousness, and Openness appear to be potent predictors of dementia and MCI, the specific facet traits within these domains appear to be differentially related to the outcome (Manning, Chan, & Stephens, in press; Terracciano et al., 2013; Williams, Suchy & Kraybill, 2013). Even more recently, arguments for the assessment of differential item relations within specific facet scales have emerged (Mottus, 2016). On the NEO-FFI, at least 1 item is present from 26 of the 30 NEO-PI R facet scales (NEO-PI R facets with no representation included N5 (Impulsiveness), A5 (Modesty), C1 (Competence), and C6 (Deliberation)). This collection of 26 items or item parcels was used in order to afford the specificity of single items (or small item clusters), while retaining a theoretical mapping to the constructs of the NEO-PI R facets. Supplement Table S2 lists the items / parcels corresponding to NEO-PI R facets. All items and item parcels were centered on the five-point Likert scale midpoint so that zero reflected an average Likert response of Neutral, with one point reflecting the shift between two Likert categories (e.g., agree to strongly agree) for the item or average item in a parcel.

Analyses

Cox Proportional Hazard models were employed, as the outcome was event occurrence at a particular follow-up point in the presence of censoring. All predictors were screened prior to analysis for proportional hazard violations and non-linearity in the log hazard. Separate risk scores were developed from prediction models based on a) only demographic factors and b) demographic factors and candidate health risk factors, and c) model b + personality items / parcels. A risk score of this type indicates the log of the individual’s hazard, or probability of experiencing the event, relative to a theoretical person scoring zero on all risk factors in the model[4]. An initial demographic-only risk score was estimated based strictly on the three demographic factors, with weights shrunk by ridge-penalized Cox models (see Hastie et al. 2009, or Chapman et al., 2016 for discussions of penalization-based shrinkage estimation to prevent overfitting). A second demographic/medical risk score was then developed by an elastic-net penalized Cox model, which selected the most predictive health factors, controlling for demographics. Weights for these health factors and demographics were then estimated with ridge-penalized cox models. A similar approach was taken for a third risk score adding personality items/parcels (selection with elastic net models, controlling for demographics and previously selected health risks, followed by weight estimation based on ridge-penalized Cox models). Tuning parameters governing selection and shrinkage were

[4] Taking the exponent of the linear predictor provides the relative hazard, with values above 1 indicated increased probability, relative to a theoretically risk-free person, and values below 1 indicating reduced probability of the outcome.

determined here by jack-knife cross-validation, as those resulting in the least shrinkage needed to achieve cross-validation error within one standard error of the minimum achievable cross-validation error[5]. Risk scores were then estimated from each of these three models using leave-one out cross-validation to afford some approximation to out-of-sample prediction. The degree of shrinkage was evaluated later by calibration slopes in ML-estimated Cox models; values over one would indicate some degree of protection against overfitting.

Performance was assessed by two overall measures, Nagelkerke's pseudo R^2 and the Brier score (Steyerberg et al., 2010). The former was based on Royston's (2006) adaptation for survival models (here abbreviated R^2_{NR}). Discrimination was assessed by the AUC, supplemented with Royston's (2006) discrimination R^2, or R^2_D, which reflects explained variation in survival outcomes. All of these measures range from zero to one, with one reflecting better model performance. Calibration was assessed by calibration in the large (the deviation between model predicted average incidence and actual incidence of the outcome), and the Hosmer-Lemeshow (HL) statistic (over- or under- prediction of the outcome across a given range of risk scores). Higher values of each measure indicate worse calibration. Finally, we examined the number of cases and non-cases that would be correctly and incorrectly identified per 100 patients in a hypothetical screening program interested in maximizing sensitivity in general medical settings, so that possible cases could be referred to specialty clinics for more detailed assessment with imaging. That strategy in general (regardless of the prediction model used) assumes that false positives are less costly than actually missing real cases who will go on to develop MCI. However, any auxiliary data that could reduce the number of false positives at a given high sensitivity might be viewed as clinically valuable, since MRI scans and other imaging tests are often very expensive.

Results

Of 386 persons free from impairment at baseline contributing a total of 786 person-years of follow-up time, 78 individuals met possible MCI criteria over the follow-up, a 4-year incidence rate of roughly 20%. Table 1 shows the estimated weights for the demographics only model (model A). These values indicate that each year of age above 75 increases the log relative hazard of possible MCI by .057, each year of education above an associate's degree decreases it by around this much, and being female decreases it substantially more. The next column, model B, shows weights for health factors identified as the most potent risks by elastic net models. Most cerebrovascular factors and burden in some organ systems increase risk. A few decrease risk, possibly because they proxy helpful treatment or increased medical attention in a primary care population (for instance, high cholesterol may indicate treatment with statins). The final column shows the weights for model C, which augments model B components with the NEO-FFI items and parcels evidencing the strongest predictive power in elastic net selection. Associations are in the same direction as those reported in NEO-PI-R wide facet- analyses for full-blown dementia in a community

[5] Conventionally, the most shrinkage producing cross-validation error within one SE of the minimum is used (see Hastie et al., 2009). This informal rule was developed for classification and regression trees, and can induce extreme underfitting of the data. Our approach is anti-conservative compared to this, seeking to achieve modest protection against overfitting in an initial model-development application, without severe underfitting.

sample (Terracciano et al., 2013), as well as Openness (Williams et al., 2013) and Neuroticism (Manning et al., 2017) facet analysis for lesser degrees of cognitive impairment in other populations. The positive emotions item parcel is an exception, with an average shift of one Likert scale point corresponding to a log hazard increase of .30. Calibration slopes for the risks scores from cross-validated predictions were all over 1, and indicated that the weights in Table 2 were on average were 25% to 35% smaller than their MLEs. Thus, the selection and shrinkage procedures reduced MLE overfitting to some degree.

Figure 1, panels A-C show the distribution of the risk scores from these models. The values of the scores represent log relative hazards of possible MCI over the four-year follow-up, compared to a hypothetical person with zero on all factors (i.e., for model C, a 75 year old man with an associate's degree, perfect health, and responding "neutral" to all personality items). Graphically, model A shows potential bimodality. Adding health risk factors affords more variability compared to demographics only, with the addition of personality items and parcels increasing variability and normalizing score distributions still more. Separation between case and control distributions also appears to increase across the three models, with Cohen's D for differences in risk scores being .5 (95% Confidence Interval [CI]) = .29, .80) for model A, .68 (95% CI = .40, .90) for model B, and .84 (.60, 1.08) for model C.

Table 2 shows the performance measures for each set of risk scores. In terms of overall performance, the R^2_{NR} is improved by adding medical factors to demographics, but not significantly so (p = .12, bootstrapping standard errors for this and all other differences). The scaled Brier score is actually worse when biomedical factors are added (p = .001). Adding the NEO items / parcels in model C results in significant improvement in overall performance, compared to the demographics and medical risk model (p = .009 for R^2_{NR} and p < .001 for the scaled Brier).

A similar pattern is apparent for the primary measure of discrimination, the AUC: non-significant improvements are achieved by the medical risk factors over the demographics only risk score, with larger and significant increments when NEO items are added (p < .001). Figure 2 shows the AUC for the three risk scores. Model C, incorporating NEO items, shows gains in sensitivity particularly over false positive rates ranging from about .2 to .5. A sensitivity of .75 is achieved at a false positive rate of around .35 for the full model, with false positive rates of .45 to.5 needed to reach a .75 sensitivity in models A and B. The supplementary R^2_D mirrors this pattern, showing a trend toward improved discrimination for model B over A (p = .09) and larger, significant improvements when personality measurement is added in C vs. B (p = .001). Royston (2009) reports that values of .4 on this measure are typical in prognostic models in oncology settings.

For calibration in the large, Model A scores show an average probability of about .10 higher than the actual outcome prevalence of .2 over the 4 year follow-up period. Model B over predicts the outcome rate significantly more (p < .001), being .15 too high on average of suggesting about 15% too many cases. Evidence of this can be seen in the score distributions of Figure 1, which are shifted slightly to the right on the x-axis. Adding NEO items results in an average probability very near the actual outcome incidence, as indicated by a deviation of near zero in Table 2 (a substantial improvement over Model B, p < .001). A similar picture appears in the HL statistic, reflecting misfit. Adding medical factors alone to demographics renders misfit worse (a higher HL value), owing to disproportionate over prediction in the highest two quintiles of risk scores. Model C, however, shows a greatly reduced and non-significant HL value (p = .742), suggesting that NEO-items have mitigated the systematic over prediction.

Under a clinical scenario in which the cut-point for deciding potential cases is set at a sensitivity of .90, and assuming the observed 4-year incidence of possible MCI of .20, the biomedical risk score, score B, predicts about 75 cases per 100 patients. Of the 20 actual cases per 100 patients at this incident rate, 18 are captured by a cut-point with .90 sensitivity. That same cut-point produces a high number of false positives as expected, in this case declaring 57 out of 80 normal patients as cases. The addition of NEO items in Model C, maintaining a cut-point with .90 sensitivity to capture the same 18 out of 20 true cases, results in a 53 false positives, with 4 error shifted to correct negative predictions. This is a single scenario based on one cut-point strategy and an assumed base rate empirically determined from the sample. Other scenarios lead to differing shifts across categories, and can generally be determined using the false positive rate for a given sensitivity in Figure 1 and an assumed base rate. Still other approaches weight false negatives and false positives differently, depending on the relative costs of mistakes (Steyerberg, 2010).

Summary and Discussion of Illustration

In this section, we presented a possible implementation of personality data in medical risk prediction models. We examined whether NEO-FFI items could improve upon a standard health and demographic risk score, for 4-year incidence of possible MCI among older persons in primary care clinics. Results indicated modest improvements overall, which could be broken down into both discrimination improvement and more substantial calibration improvement. To take one possible implementation scenario, these risk scores would be computed in a general medical setting as a first line prediction erring on the side of sensitivity, and referring those testing positive to specialty clinics that could collect more costly and detailed data, such as MRI scans. Using a cut-point in all risk scores that would capture 90% of the true cases, with the 20% incidence rate in this cohort, NEO items move false positives from 57 / 80 (71%) patients who will not develop possible MCI to 53 / 80 (66%), or reduce about four unneeded MRIs in 57, a significant potential cost savings. In general, dementia risk prediction models tend to have AUCs in the .7's in the samples in which they are developed (and without protection from overfitting; Stephan et al., 2010), a figure approximated by the personality-integrated model. From the standpoint of the criteria we outlined, this application of personality data to prediction models would appear modestly successful as a starting point. However, standards for clinical use are far higher (for instance, AUCs in the .9s), so the results would represent only an initial foundation in need of considerably more development. In future development of personality-augmented predictive models in this specific example area, several other lines of work are also needed. These include whether personality change prior to cognitive decline is a predictor itself or compromises the predictive value of self-reports, whether informant reports of personality might be helpful, and whether item (vs. facet) level data represent better levels of analysis.

In the following section, we take a contrarian perspective, playing devil's advocate in an effort to highlight the resistance that actual implementations such as this may face. Our goal is not undue negativity, but merely to point out the real-world challenges that exist in attempting to implement an idea from one area (personality psychology) in a different context (health care).

Challenges Moving Forward

Stakeholder Attitudes

Healthcare Provider Attitudes Toward Risk Models—Physicians do and will likely continue to have varying attitudes toward the utility of risk prediction models. Thus, before even considering whether or when personality may be of use, a fundamental question is to what extent existing models are used, and in what cases, and how they inform clinical decision-making. Answers will likely vary drastically across different contexts. Some classic risk models, such as the Framingham family of cardiovascular outcome models have a substantial evidence base and thus appear to inform clinical practice at least somewhat. In other areas, the accuracy of risk models lag far behind standards for clinical practice. These are precisely the areas where the most work is needed. The prospects for personality augmentation are thus probably best at the nexus of outcomes for which personality has demonstrated predictive power, and for which existing models are not already strong.

Biomedical Biases—It is quite possible that even if a personality item index works as well or better than a polygenic risk score, personality measurement may be deemed unacceptable by some physicians. Some within health care have been trained within a heavily biomedical tradition. Even in those trained or practicing in a biopsychosocial model, implicit biomedical bias is powerful and pervasive. Such bias can range from beliefs that personality cannot be studied scientifically, to assumptions that it cannot be measured with any degree of accuracy, to incredulity that it could hold any relevance for a given health outcome. Appeals to the biological bases of personality occasionally remediate these biases, although they can also create the impressions that a trait measurement is merely a stand-in for some genotype. Other physicians are quite open-minded or even in some cases actively aware of personality research in their field. There have also been secular shifts in attitudes toward non-biological determinants of health. Recent changes to medical school curricula now duly acknowledge behavioral and social determinants of health (American Association of Medical Colleges, 2011), which will likely result in greater receptivity among younger physicians and in coming years. The key point is not that biomedical bias is all-pervasive and insurmountable, but rather that it may at times be encountered.

Medical administrators and management represent another set of key stakeholders. These individuals are business people focused on maximizing the profit of a practice or health care system. If it could be demonstrated that any auxiliary risk prediction measurements ultimately improve outcomes and reduces costs, such persons might be supportive of implementation. In the example above, if the procedure is to utilize a highly sensitive cut-point and refer those testing "positive" for expensive imaging procedures, reducing 4 unneeded MRIs in 100 might be deemed noteworthy cost-savings. On the other hand, if providers and or patients bridled so much at the idea that business was lost and bottoms lines damaged, the chances of any real implementation are greatly reduced. It would be helpful, moving forward, to examine issues around provider and administrator attitudes toward any efforts to concretely implement personality measurement within real-world health care settings.

Patient Attitudes—Patients themselves may or may not accept the administration of personality inventory when they arrive for a medical appointment. Acceptance is likely to depend in part on the presence of a face-valid rationale. For instance, an individual arriving for a mental health appointment would likely view personality assessment as a plausible or justified request for psychological information. A self-regulation questionnaire presented at a weight loss clinic might similarly appear logical to patients. A patient arriving for a cardiology consult asked to fill out a hostility questionnaire might cooperate, given an explanation that this trait has been linked to cardiovascular outcomes. But a patient arriving at a primary care appointment for an annual check-up, confronted with a 300-item trait questionnaire without a clear rationale provided, might perceive the request as irrelevant, invasive, or both. Willingness to provide personality data might even be linked to trait levels themselves. A disagreeable person might oppose the request, while one prone to anxiety might worry about how the information will be perceived or whether it will be kept confidential. If any traits affecting measurement refusal are also important in predicting the outcome, the resulting prediction models will be compromised. On the other hand, some--if not many--individuals are heavily influenced by the trappings of medical authority, and would likely comply with requests for personality data. Yet another concern is whether valid information would be provided on a questionnaire viewed in the same category as health history, and thus subject to evaluation by unknown authority figures such as doctors or insurance companies. Patient response and acceptability therefore requires careful consideration in efforts to integrate personality into health improvement. One line of work that may provide some guidance involves the evolution of anxiety and depression screening in primary care. Questionnaires measuring those constructs progressed from unwelcome assessments of stigmatized constructs to relatively common pieces of health information in many health systems, thanks to a focus on implementation research.

Logistics—Even if providers, management, and patients are amenable to completing personality measures in a given health care clinic, there is still the matter of how the data should be collected, stored, and used. Paper-and-pencil depression and anxiety screens that staff can hand-score quickly and enter into EMRs have been in use for some time. The Patient-Reported Outcomes Measurement Information System (PROMIS; Cella et al., 2010) and similar initiatives of the NIH have been developed with computerized data-collection and linkage to EMRs in mind. From a technological standpoint, it would not be difficult in principle to administer a personality measure within a waiting room or remotely via an online patient portal.

The question of what personality items or scales to collect is an entirely different matter. The Grid Enabled Measures initiative of the NIH (https://www.gem-beta.org/Public/Home.aspx) is an effort to delineate the most useful psychosocial measures for healthcare settings. A specialty clinic might be interested only in a focal set of items related to their outcomes of interest. In the above example, for instance, administering only the 13 items showing incremental validity over biomedical risk factors would be expedient, but would prevent the calculation of descriptive scores on the Big Five domains themselves. Administering an entire instrument such as the NEO FFI has the appeal of being able to provide a general description of the patient, in addition to yielding a larger item pool from which to construct predictive models for several different health outcomes.

The downside to this, of course, is that instrument length becomes a key factor in the feasibility of real-world implementation. Even nine-item depression measures like the Patient Health Questionnaire (PHQ-9) have been deemed “long” and reduced to 2 item screeners (Kroenke, Spitzer, & Williams, 2003). The PROMIS platform offers computerized adaptive testing that can provide valid scores for an entire scale based on response patterns to key subsets of items. Most psychologists, however, know that there is “no free lunch” in psychometrics, particularly when the success of personality-augmented predictive model program likely depends on a wider item pool. Standards in the length of time considered acceptable for measurement vary between psychometrics and biometrics. A set of over 20,000 SNPs is far quicker to collect than responses to 20,000 personality items. On the other hand, the single blood draw required for genetic data or other biomarkers requires the patient to go to a clinic, wait in line, and receive a venipuncture. The total time and effort of completing the NEO-FFI’s 60 self-report items may be less. Another consideration is that a personality inventory collected for health prediction purposes would typically be collected once when a new patient enters a health system, while biomarkers, PROMIS scales and depression and anxiety screens are administered repeatedly because they are considered outcomes that need to be tracked. That is not to argue that personality does not change, but merely to draw a distinction between the likely real-world frequency with which traits would be measured and that of other measurements common within health care. Whatever one’s perspective on these issues, the logistics of data collection require careful consideration if personality measurement in healthcare is ever to reach implementation.

Third Party Abuses—We assume that information collected will be stored in EMRs, which can simply apply formulas for risk-scores and provide them as needed to health care personnel. While EMRs offer great potential, larger concerns revolve around misuse of the personality data by private and public parties who access health data. Despite the advent of the Health Information Portability and Accountability Act (HIPAA) of 1996, these entities can easily obtain sensitive health data. Insurers—be they government agencies like Medicaid and Medicare, or private companies--receive data directly from health care visits and make benefits contingent on them. This opens up the possibility of insurance discrimination, particularly among private insurers, against patients evidencing high risk on personality measures. Concerns over insurance discrimination already exist for some high-profile genetic tests. For example, those who pursue testing for the APOE polymorphism implicated in late-onset Alzheimer’s Disease are advised to purchase long-term care or other insurance policies prior to testing (Rahman et al., 2012). It is not difficult to envision “pre-existing conditions” clauses being stretched to low Conscientiousness scores to limit or deny coverage. Private insurers will not care whether academics or ethicists would consider such a decision justified if it is a profitable one.

Another concern arises out of extending mental health stigma to those with socially undesirable trait profiles. While personality trait data is fundamentally different than a mental health diagnosis, a number of people capable of accessing health records do not have the training to make such a distinction. The line between a formal "personality disorder" and a personality trait is important to the readership of this journal, but to regrettably few others. The outcome of life goals and events ranging from adoption, to child custody cases, to civil proceedings, to firearm purchases, to employment and disability programs can hinge on psychological data in medical records. Attorneys, policy makers, and bureaucrats are often third party consumers of medical records data, but ill-suited for rendering important judgments about their contents. Efforts to integrate personality into healthcare will therefore need to consider how to prevent unintended consequences from the inclusion of such data in EMRs.

Conclusion

In this paper, we have laid out the rationale for an alternative use of personality information to improve health. Rather than attempt to modify health-damaging traits directly, it is also possible to leverage the long-touted predictive power of personality to improve the formal health-prediction models that are proliferating within medicine. We then illustrated how this could be done, in an example predicting the emergence of possible MCI over 4 years within asymptomatic older primary care patients. A risk model based on demographic and biomedical factors was improved by 13 personality items or item-clusters. We then concluded by discussing the challenges that lay ahead in efforts to include personality measurement within health care. These involve patient, physician, and health system leadership reactions to the notion of assessing personality in healthcare contexts. To the extent that demonstrable improvements in the prediction of important outcomes can be wedded to low costs, various stakeholders may be more or less amenable to the idea. The potential for misuse and abuse of personality measurement by outside entities also exists, just as it does for high-risk genotypes or other sensitive health data maintained in EMRs. Moving forward, each of these issues requires considerable and systematic attention if the literature on personality and health is to realize practical impact in everyday health care settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support:

National Institutes of Health R01MH61429 (JML), R01AG044588 (BPC)

References

Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Petersen RC. 2011; The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia. 7(3):270–279.

Bogg T, Roberts BW. 2004; Conscientiousness and health-related behaviors: a meta-analysis of the leading behavioral contributors to mortality. Psychological bulletin. 130(6):887. [PubMed: 15535742]

Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Hays R. 2010; The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. Journal of clinical epidemiology. 63(11): 1179–1194. [PubMed: 20685078]

Chapman BP, Hampson S, Clarkson J. 2014; Personality Intervention for Healthy Aging: Conclusions from a National Institute on Aging Workgroup. Developmental psychology. 50(5):1411–1426.

Chapman BP, Roberts B, Duberstein P. 2011; Personality and longevity: knowns, unknowns, and implications for public health and personalized medicine. Journal of Aging Research. 2011:24.doi: 10.4061/2011/759170

Chapman BP, Weiss A, Duberstein PR. 2016; Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. Psychological Methods. 21(4):603–620. [PubMed: 27454257]

Chapman BP, Weiss A, Fiscella K, Muennig P, Kawachi I, Duberstein P. 2015; Mortality Risk Prediction: Can Comorbidity Indices Be Improved With Psychosocial Data? Medical Care. 53(11): 909–915. [PubMed: 26421372]

Christidi F, Kararizou E, Triantafyllou N, Anagnostouli M, Zalonis I. 2015; Derived Trail Making Test indices: demographics and cognitive background variables across the adult life span. Aging, Neuropsychology, and Cognition. 22(6):667–678.

Colleges, A. A. o. M.. Behavioral and Social Science Foundations for Future Physicians. 2011. Retrieved from Washington DC:

Control, C. f. D.. Health, United States, 2015. 2016. Retrieved from

Cook SE, Marsiske M, Thomas KR, Unverzagt FW, Wadley VG, Langbaum JB, Crowe M. 2013; Identification of mild cognitive impairment in ACTIVE: algorithmic classification and stability. Journal of the International Neuropsychological Society. 19(01):73–87. [PubMed: 23095218]

Deckers K, Boxtel MP, Schiepers OJ, Vugt M, Muñoz Sánchez JL, Anstey KJ, Kivipelto M. 2015; Target risk factors for dementia prevention: a systematic review and Delphi consensus study on the evidence from observational studies. International journal of geriatric psychiatry. 30(3):234–246. [PubMed: 25504093]

Drane DL, Yuspeh RL, Huthwaite JS, Klingler LK. 2002; Demographic characteristics and normative observations for derived-trail making test indices. Cognitive and Behavioral Neurology. 15(1):39–43.

Folstein MF, Folstein SE, McHugh PR. 1975; "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. Journal of psychiatric research. 12(3):189–198. [PubMed: 1202204]

Goff, DC; Lloyd-Jones, DM; Bennet, G; Coday, S; D'Agostino, RB; Gibbons, R; , et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk. Circulation. 2013. Novermber 12 2013, online issue

Gordis, L. Epidemiology. Third. Philadelphia, Pennsylvania: Saunders; 2004.

Greenland S. 2008; The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI:10.1002/sim. 2929). Statistics in medicine. 27(2):199–206. DOI: 10.1002/sim.2995 [PubMed: 17729377]

Hamburg MA, Collins FS. 2010; The path to personalized medicine. The New England journal of medicine. 363(4):301–304. DOI: 10.1056/NEJMp1006304 [PubMed: 20551152]

Hastie, T, Tibshirani, R, Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. New York: Springer; 2009.

Heister D, Brewer JB, Magda S, Blennow K, McEvoy LK. Initiative, A. s. D. N. 2011; Predicting MCI outcome with clinically available MRI and CSF biomarkers. Neurology. 77(17):1619–1628. [PubMed: 21998317]

Hester RL, Kinsella GJ, Ong B, McGregor J. 2005; Demographic influences on baseline and derived scores from the trail making test in healthy older Australian adults. The Clinical Neuropsychologist. 19(1):45–54. [PubMed: 15814477]

Institute of, M.. Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. 2006. Retrieved from Washington, DC:

James BD, Leurgans SE, Hebert LE, Scherr PA, Yaffe K, Bennett DA. 2014; Contribution of Alzheimer disease to mortality in the United States. Neurology. 82(12):1045–1050. [PubMed: 24598707]

Kaffashian S, Dugravot A, Elbaz A, Shipley MJ, Sabia S, Kivimäki M, Singh-Manoux A. 2013; Predicting cognitive decline A dementia risk score vs the Framingham vascular risk scores. Neurology. 80(14):1300–1306. [PubMed: 23547265]

Kantarci K, Weigand SD, Przybelski SA, Preboske GM, Pankratz VS, Vemuri P, Machulda MM. 2013; MRI and MRS predictors of mild cognitive impairment in a population-based sample. Neurology. 81(2):126–133. [PubMed: 23761624]

Kroenke K, Spitzer RL, Williams JB. 2003; The Patient Health Questionnaire-2: validity of a two-item depression screener. Medical Care. 41(11):1284–1292. [PubMed: 14583691]

Lamberty, GJ, Axelrod, BN. 10 Derived adult Trail Making Test indices. In: Porah, A, editorThe Quantified Process Approach to Neuropsychological Assessment. Taylor & Francis; 2006. 161

Linn BS, Linn MW, Gurel L. 1968; Cumulative illness rating scale. Journal of the American Geriatrics Society. 16(5):622–626. [PubMed: 5646906]

Low L-F, Harrison F, Lackersteen SM. 2013; Does personality affect risk for dementia? A systematic review and meta-analysis. The American Journal of Geriatric Psychiatry. 21(8):713–728. [PubMed: 23567438]

Lucas JA, Ivnik RJ, Smith GE, Bohac DL, Tangalos EG, Kokmen E, Petersen RC. 1998; Normative data for the Mattis dementia rating scale. Journal of clinical and experimental neuropsychology. 20(4):536–547. [PubMed: 9892057]

Luchetti M, Terracciano A, Stephan Y, Sutin AR. 2016; Personality and cognitive decline in older adults: Data from a longitudinal sample and meta-analysis. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences. 71(4):591–601.

Manning KJ, Chan G, Stephens DC. Neuroticism Traits Selectively Impact Long Term Illness Course and Cognitive Decline in Late-Life Depression. American Journal of Geriatric Psychiatry.

Mroczek DK. 2014; Personality plasticity, healthy aging, and interventions. Developmental psychology. 50(5)

O’Bryant SE, Humphreys JD, Smith GE, Ivnik RJ, Graff-Radford NR, Petersen RC, Lucas JA. 2008; Detecting dementia with the mini-mental state examination in highly educated individuals. Archives of neurology. 65(7):963–967. [PubMed: 18625866]

Pankratz VS, Roberts RO, Mielke MM, Knopman DS, Jack CR, Geda YE, Petersen RC. 2015; Predicting the risk of mild cognitive impairment in the Mayo Clinic Study of Aging. Neurology. 84(14):1433–1442. [PubMed: 25788555]

Pepe MS, Fan J, Feng Z, Gerds T, Hilden J. 2015; The net reclassification index (NRI): a misleading measure of prediction improvement even with independent test data sets. Statistics in biosciences. 7(2):282–295. [PubMed: 26504496]

Peterson RE, Maes HH, Holmans P, Sanders AR, Levinson DF, Shi J, Webb BT. 2011; Genetic risk sum score comprised of common polygenic variation is associated with body mass index. Human genetics. 129(2):221–230. [PubMed: 21104096]

Piantidosi, S. Clinical Trials: A Methodologic Perspective. Second. New York: John Wiley & Sons; 2005.

Rahman B, Meiser B, Sachdev P, Barlow-Stewart K, Otlowski M, Zilliacus E, Schofield P. 2012; To know or not to know: an update of the literature on the psychological and behavioral impact of genetic testing for Alzheimer disease risk. Genetic testing and molecular biomarkers. 16(8):935–942. [PubMed: 22731638]

Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HC. 2007; Prognosis research: toward evidence-based results and a Cochrane methods group. Journal of clinical epidemiology. 60(8): 863–865. [PubMed: 17606185]

Roberts BW, Kuncel N, Shiner RN, Caspi A, Goldberg L. 2007; The power of personality: A comparative analysis of the predictive validity of personality traits, SES, and IQ. Perspectives in Psychological Science. 4(2):313–346. DOI: 10.1111/j.1745-6916.2007.00047

Rost K, Smith J. 2001; Retooling multiple levels to improve primary care depression treatment. Journal of general internal medicine. 16(9):644–645. [PubMed: 11556948]

Royston P. 2006; Explained variation for survival models. Stata Journal. 6(1):83–96.

Royston P, Moons KG, Altman DG, Vergouwe Y. 2009; Prognosis and prognostic research: Developing a prognostic model. BMJ. 338:b604. [PubMed: 19336487]

Schlendorf KH, Nasir K, Blumenthal RS. 2009; Limitations of the Framingham risk score are now much clearer. Preventive medicine. 48(2):115–116. DOI: 10.1016/j.ypmed.2008.12.002 [PubMed: 19124038]

Sperling RA, Karlawish J, Johnson KA. 2013; Preclinical Alzheimer disease—the challenges ahead. Nature Reviews Neurology. 9(1):54–58. [PubMed: 23183885]

Stephan BC, Kurth T, Matthews FE, Brayne C, Dufouil C. 2010; Dementia risk prediction in the population: are screening models accurate? Nature Reviews Neurology. 6(6):318–326. [PubMed: 20498679]

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Kattan MW. 2010; Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 21(1):128–138. DOI: 10.1097/EDE.0b013e3181c30fb2 [PubMed: 20010215]

Terracciano A, Iacono D, O'Brien RJ, Troncoso JC, An Y, Sutin AR, Resnick SM. 2013; Personality and resilience to Alzheimer's disease neuropathology: a prospective autopsy study. Neurobiology of aging. 34(4):1045–1050. [PubMed: 23040035]

Tzoulaki I, Liberopoulos G, Ioannidis JP. 2009; Assessment of claims of improved prediction beyond the Framingham risk score. JAMA : the journal of the American Medical Association. 302(21): 2345–2352. [PubMed: 19952321]

Unverzagt F, McClure L, Wadley V, Jenny N, Go R, Cushman M, Moy C. 2011; Vascular risk factors and cognitive impairment in a stroke-free cohort. Neurology. 77(19):1729–1736. [PubMed: 22067959]

Weiss A, Gale CR, Batty GD, Deary IJ. 2013; A questionnaire-wide association study of personality and mortality: The Vietnam Experience Study. Journal of psychosomatic research. 74(6):523–529. [PubMed: 23731751]

Williams PG, Suchy Y, Kraybill ML. 2013; Preliminary evidence for low openness to experience as a pre-clinical marker of incipient cognitive decline in older adults. Journal of Research in Personality. 47(6):945–951.

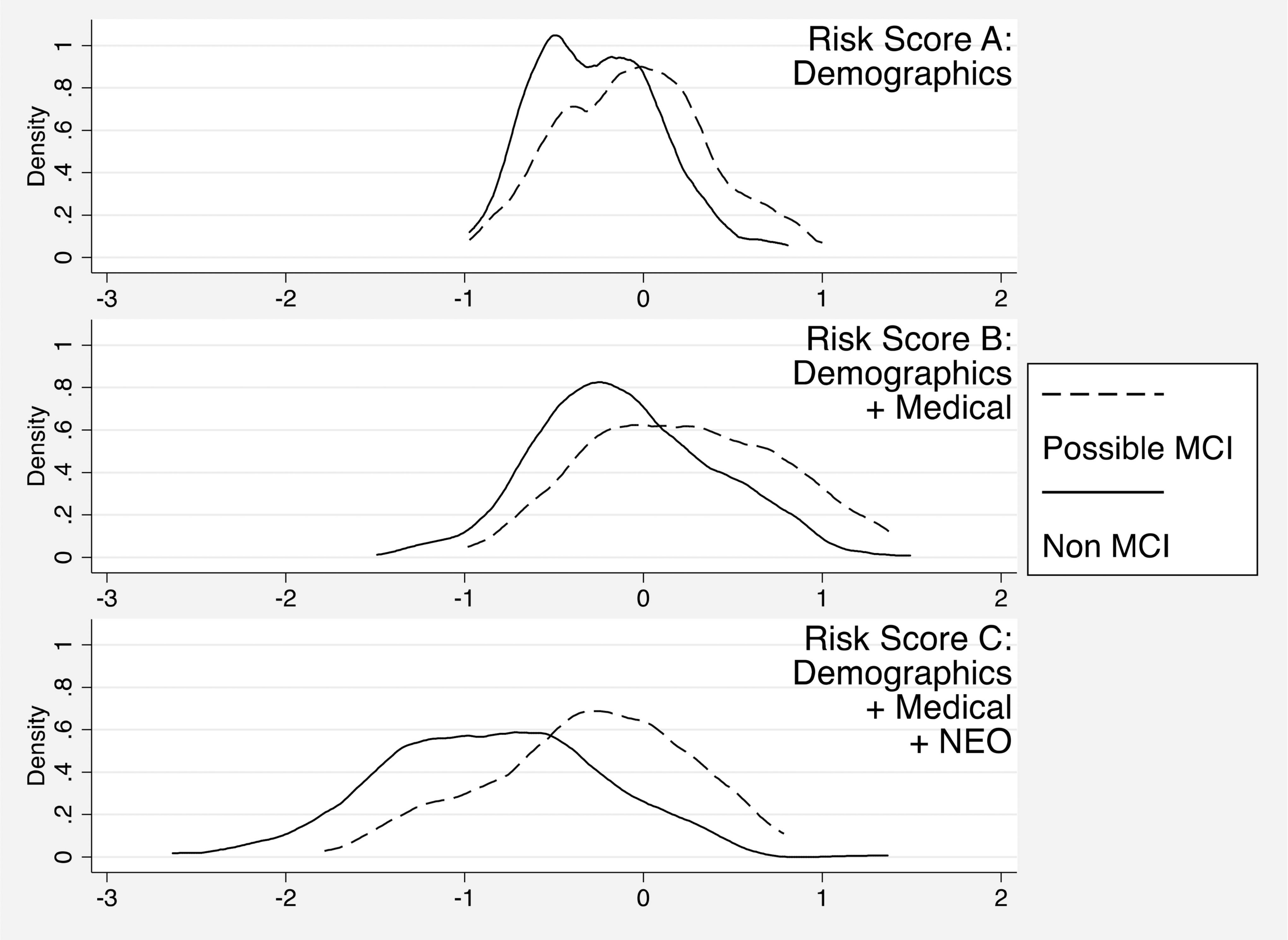

Figure 1.

Distribution of risk scores from three prediction models of Mild Cognitive Impairment. Log relative hazard metric. Model A = demographics only, Model B = Model A + health factors, Model C = Model B + personality items/ facets.

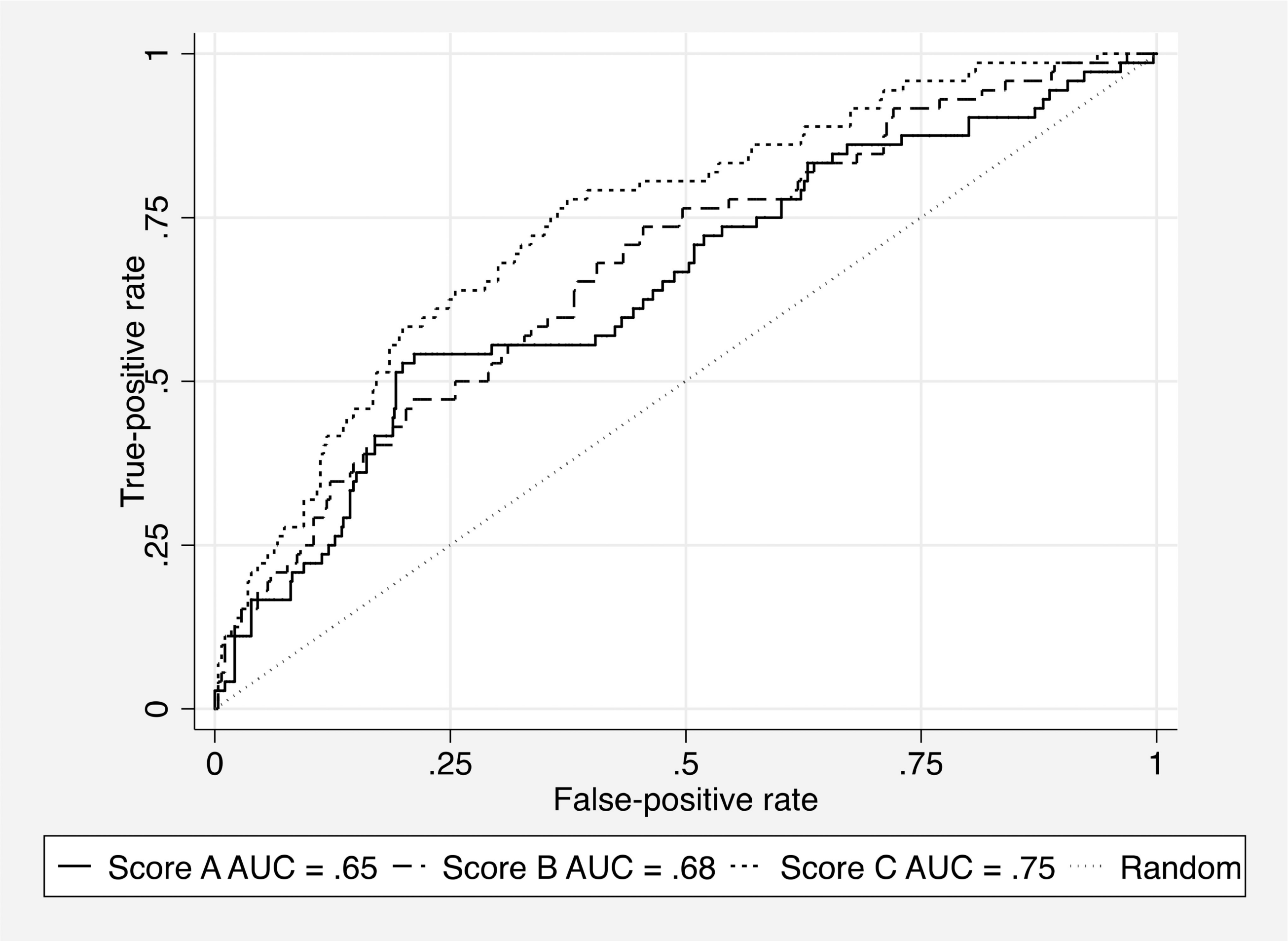

Figure 2.

Area under curve for three prediction models of Mild Cognitive Impairment. Model A = demographics only, Model B = Model A + health factors, Model C = Model B + personality items/ facets.

Table 1

Weights for Risk Models of Four-Year MCI Incidence

	A	B	C
Demographics			
Age	0.057	0.055	0.05
Female	−0.256	−0.244	−0.16
Education	−0.050	−0.053	−0.03
Health Factors			
Diabetes		0.146	0.11
Current Smoker		0.548	0.48
Cardiovascular Disease		0.173	0.25
Atrial Fibrillation		0.342	0.36
Left Ventricular Hypertrophy		0.144	0.24
Former Smoker		0.053	0.07
Total Cholesterol		−0.103	−0.11
HDL Cholesterol		−0.013	−0.01
CIRS Respiratory		0.032	0.00
CIRS Urinogenial		0.156	0.10
CIRS Gastrointestinal		0.244	0.21
CIRS Neurologic		−0.063	−0.06
CIRS Endocrine		−0.080	−0.10
Personality Item/Parcels			
N1 Anxiety			0.18
N2 Angry Hostility			0.06
E1 Warmth			−0.15
E6 Positive Emotion			0.30
O2 Aesthetics			−0.22
O3 Emotions			−0.13
A6 Tendermindedness			−0.05
C3 Dutifulness			−0.23
C5 Self-discipline			−0.18

Notes: N = 358, log hazard rates from regularized Cox models. HDL = High Density Lipoprotein, CIRS = Cumulative Illness Rating Scale. Items or parcels from NEO-Five Factor Inventory corresponding to facets from NEO-Personality Inventory, Revised; N = Neuroticism, E = Extraversion, O = Openness, A = Agreeableness, and C = Conscientiousness. Model A = demographics only, Model B = Model A + health factors, Model C = Model B + personality items/ facets.

Table 2

Performance of Three Risk Models of Four-Year MCI Incidence

	A	B	C
Overall Fit			
Nagelkerke / Royston R^2	.164 (.053, .297)	.247 (.113, .384)	.374 (.226, .502)
Scaled Brier Score	.278 (.239, .313)	.226 (.164, .280)	.375 (.326, .417)
Discrimination			
AUC	.647 (.576, .718)	.680 (612,.748)	.741 (.677, .805)
Discrimination R^2	.164 (.053, .297)	.247 (.113, .384)	.374 (.226, .502)
Calibration			
Calibration in the Large	.101 (.090, .112)	.155 (.040, .171)	.002 (−.013, .01)
Hosmer-Lemeshow Statistic	18.90 ($p = .0003$)	41.89 ($p < .0001$)	3.11 ($p = .742$)

Notes: Measures of model performance with 95% Confidence Intervals in parentheses for Table 2 models. AUC = Area Under the Curve. Model A = demographics only, Model B = Model A + health factors, Model C = Model B + personality items/ facets.