Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(5) 1427–1438

DOI: 10.1177/0962280218759138
journals.sagepub.com/home/smm
SAGE

The informative error: A framework for the construction of individualized phenotypes

Johannes Hertel,[1,2] Stefan Frenzel,[1] Johanna König,[1] Katharina Wittfeld,[2] Georg Fuellen,[3] Birte Holtfreter,[4] Maik Pietzner,[5,6] Nele Friedrich,[5,6,7] Matthias Nauck,[5,6] Henry Völzke,[6,8] Thomas Kocher[4] and Hans J Grabe[1,2]

Abstract

For the goal of individualized medicine, it is critical to have clinical phenotypes at hand which represent the individual pathophysiology. However, for most of the utilized phenotypes, two individuals with the same phenotype assignment may differ strongly in their underlying biological traits. In this paper, we propose a definition for individualization and a corresponding statistical operationalization, delivering thereby a statistical framework in which the usefulness of a variable in the meaningful differentiation of individuals with the same phenotype can be assessed. Based on this framework, we develop a statistical workflow to derive individualized phenotypes, demonstrating that under specific statistical constraints the prediction error of prediction scores contains information about hidden biological traits not represented in the modeled phenotype of interest, allowing thereby internal differentiation of individuals with the same assigned phenotypic manifestation. We applied our procedure to data of the population-based *Study of Health in Pomerania* to construct a refined definition of obesity, demonstrating the utility of the definition in prospective survival analyses. Summarizing, we propose a framework for the individualization of phenotypes aiding personalized medicine by shifting the focus in the assessment of prediction models from the model fit to the informational content of the prediction error.

Keywords

Individualization, personalized medicine, prediction modelling, measurement error, directed acyclic graphs, obesity, individualized medicine

1 Introduction

The classificatory systems and thus our nomenclatures of health and disease like the ICD-10 are not designed to deliver the full pathophysiological picture of an individual patient.[1–3] The same is true for a broad range of classical risk factors like age or the body mass index (BMI) which can be seen as proxies for underlying pathophysiological processes, but they can clearly not be identified with the underlying biology.[4–6] From a conceptual point of view, it is therefore the difference between our phenotypes (mostly nomenclatures and risk factors) and the actual pathophysiology ongoing in a patient which represents a major obstacle in the development of individualized metrics describing health and disease. For the goal of personalized medicine, statistical modeling

[1]Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Germany
[2]German Center for Neurodegenerative Diseases (DZNE), Site Rostock/Greifswald, Germany
[3]Institute for Biostatistics and Informatics in Medicine and Ageing Research, Rostock University Medical Center, Rostock, Germany
[4]Unit of Periodontology, Department of Restorative Dentistry, Periodontology, Endodontology, and Preventive and Pediatric Dentistry, Dental School, University Medicine Greifswald, Germany
[5]Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Germany
[6]DZHK (German Center for Cardiovascular Research), partner site Greifswald, Germany
[7]Research Centre for Prevention and Health, Glostrup University Hospital, Denmark
[8]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

Corresponding author:
Johannes Hertel, Department of Psychiatry, University Medicine Greifswald, Ellernholzstrasse 1-2, Greifswald 17475, Germany.
Email: hertelj@uni-greifswald.de

should therefore aim at delivering metrics fitting to the individual biology instead of delivering the best fit to an observable phenotype. Moreover, when thinking of economic pressure in modern health systems, data-driven predictors would have to prove a substantial *informational surplus* in comparison to the classical risk assessments and diagnostic procedures to be accepted and refunded.[7] For example, there would be no reason to spend money on quantifying the urine metabolome[8] or the epigenome[9] to measure biological age if the simple question "How old are you?" would deliver the same information as a potentially complicated and expensive Omics analysis.

This paper deals with the question how to generate such individualized metrics of health and disease and how to test their validity. We will present a definitional statistical framework and derived a generally applicable workflow to individualize risk factors and phenotypes, by shifting the criterion of successful modeling from the *model fit* to the *informational content of the prediction error*. Utilizing measurement error theory,[10,11] we prove that the prediction error can be used for meaningful internal differentiation of individuals showing the same observable phenotype if certain statistical prerequisites are fulfilled. We shall note that our arguments have been implicitly already used in biomedical research, especially in the construction of biological age measures, without clarifying the underlying assumptions and methodology.[8,9,12–14]

2 Theoretical background

In this part, we concretize our understanding of *individualization* and introduce the statistical concept of *an individualization instrument*. We define the concept of individualization as follows:

2.1 Definition 1: Individualization (conceptual)

Individualization is the differentiation of subjects showing the same phenotypic manifestation despite underlying biological differences such that the assigned differences correspond to the underlying biological differences.

This definition of individualization makes thus only sense conceptually if a bijection between phenotype and biology is not possible. If this would the case, the phenotype could already be called individualized. We will therefore assume in the following paragraphs that at least some aspects of the underlying biology are not observed. Before transferring the definition of individualization to statistical terms, we concretize the term *proxy phenotype*:

2.2 Definition 2: Proxy phenotype

Let $\boldsymbol{T}$, $\boldsymbol{X_1}$, $\boldsymbol{X_2}$,..., $\boldsymbol{X_I}$ be random variables with $\boldsymbol{T}$ representing a hidden trait and the $\boldsymbol{X_i}$ representing observable variables. We call $\boldsymbol{X_i}$ a proxy phenotype for $\boldsymbol{T}$ if

(i) $\boldsymbol{X_i}$ *and* $\boldsymbol{T}$ *are not statistically independent* : $\boldsymbol{X_i} \not\perp\!\!\!\perp \boldsymbol{T}$
(The observable variable $\boldsymbol{X_i}$ carries information about the hidden biological trait)
(ii) $\boldsymbol{X_i}$ *and* $\boldsymbol{T}$ *are not statistically independent given all other* $\boldsymbol{X_j}$: $\boldsymbol{X_i} \not\perp\!\!\!\perp \boldsymbol{T} | \cup_{j=1, j\neq i}^{I} \boldsymbol{X_j}$
(The observable variable $\boldsymbol{X_i}$ carries information about the hidden biological trait given all other observable variables.)

By this definition, we include genotypes into the class of proxy phenotypes as long as they are observed. In the context of individualization, this makes sense because individuals having the same genotype may differ in their biology regarding a certain hidden trait. Thus, the classical differentiation between genotype and phenotype has no meaning inside our framework. We translate now the above made definition of individualization into statistical terms.

2.3 Definition 3: Individualization (statistical)

Let $\boldsymbol{M}$, $\boldsymbol{X}$ and $\boldsymbol{T}$ be three random variables. We call the observable $\boldsymbol{M}$ *an individualization in the context of the hidden trait* $\boldsymbol{T}$ *and the corresponding proxy phenotype* $\boldsymbol{X}$ if the following attributes for $\boldsymbol{X}$,$\boldsymbol{T}$ and $\boldsymbol{M}$ are given:

(i) $\boldsymbol{M}$ *and* $\boldsymbol{T}$ *are not statistically independent* : $\boldsymbol{M} \not\perp\!\!\!\perp \boldsymbol{T}$
(The individualization carries information about the hidden biological trait)
(ii) $\boldsymbol{M}$ *and* $\boldsymbol{T}$ *are not statistically independent given* $\boldsymbol{X}$: $\boldsymbol{M} \not\perp\!\!\!\perp \boldsymbol{T} | \boldsymbol{X}$
(The individualization carries information about the hidden biological trait given the phenotype $\boldsymbol{X}$)

In the context of metric phenotypes and metric hidden traits, we call $\boldsymbol{M}$ *individualization metric*. At first sight, it might be plausible to demand additionally that the individualization $\boldsymbol{M}$ should be independent of the proxy phenotype $\boldsymbol{X}$ given the hidden trait $\boldsymbol{T}$. However, this would exclude the case in which $\boldsymbol{T}$ is a function of the individualization $\boldsymbol{M}$ and the proxy phenotype $\boldsymbol{X}$. In this case, conditioning on $\boldsymbol{T}$ could lead to statistical dependence between $\boldsymbol{M}$ and $\boldsymbol{X}$ (in terms of causal inference theory, $\boldsymbol{T}$ would be a collider[15]). Still, the individualization $\boldsymbol{M}$ would carry information regarding the latent trait $\boldsymbol{T}$ not represented in $\boldsymbol{X}$. Next, we derive conditions under which a variable $\boldsymbol{Z}$ associated to $\boldsymbol{X}$ may be *useful* in deriving individualizations, and here indeed, we will demand conditional independence between $\boldsymbol{Z}$ and $\boldsymbol{X}$ given $\boldsymbol{T}$.

2.4 Definition 4: Individualization instrument

Let $\boldsymbol{Z}$, $\boldsymbol{X}$ and $\boldsymbol{T}$ be three random variables. We call the observable $\boldsymbol{Z}$ *an individualization instrument in the context of the hidden trait* $\boldsymbol{T}$ *and the corresponding proxy phenotype* $\boldsymbol{X}$ if the following attributes for $\boldsymbol{X}$,$\boldsymbol{T}$ and $\boldsymbol{Z}$ are given:

(i) $\boldsymbol{Z}$ *and* $\boldsymbol{T}$ *are not statistically independent* : $\boldsymbol{Z \not\perp\!\!\!\perp T}$
(The individualization instrument carries information about the hidden biological trait)
(ii) $\boldsymbol{X}$ *and* $\boldsymbol{Z}$ *are statistically independent given* $\boldsymbol{T}$: $\boldsymbol{X \perp\!\!\!\perp Z|T}$
(The phenotype carries no information about the individualization instrument given the value of the hidden biological trait.)

This collection of statistical dependencies corresponds to three possible Bayesian nets which can be visualized by directed acyclic graphs:

(a) $\boldsymbol{Z \leftarrow T \rightarrow X}$
(b) $\boldsymbol{Z \rightarrow T \rightarrow X}$
(c) $\boldsymbol{Z \leftarrow T \leftarrow X}$

These directed acyclic graphs may have causal interpretations and it may be helpful to think of these relations in causal terms (in the sense of Pearl[15]), but we do not rely in our framework on the notion of causality. It follows from (i)-(ii) directly that $\boldsymbol{X \not\perp\!\!\!\perp Z}$ and $\boldsymbol{Z \not\perp\!\!\!\perp T|X}$**.** The latter means that the individualization instrument $\boldsymbol{Z}$ contains information about $\boldsymbol{T}$ in individuals showing the same phenotypic occurrence. Thus, an individualization instrument is a statistical individualization in the sense of definition 3. The converse statement is obviously not true. Consider the acyclic graph (d):

(d) $\boldsymbol{Z \rightarrow T \leftarrow X}$

In this case, $\boldsymbol{Z}$ is not an individualization instrument, because $\boldsymbol{Z}$ is not independent of $\boldsymbol{X}$ given $\boldsymbol{T}$ (once again $\boldsymbol{T}$ can be seen as a collider), but $\boldsymbol{Z}$ is clearly an individualization. The justification for the exclusion of the case (d) (and other cases of individualizations) from the definition of an *individualization instrument* is given later on when we will discuss the construction of individualized proxy phenotypes from a set of observable variables.

An example for an individualization instrument is C-reactive protein measures in serum (variable $\boldsymbol{Z}$) as a marker for chronic low-grade inflammation in the context of hidden trait biological age (variable $\boldsymbol{T}$) and the proxy phenotype chronological age (variable $\boldsymbol{X}$). The biological age variable $\boldsymbol{T}$ is a function of the chronological age variable $\boldsymbol{X}$. Chronic low grade inflammation is a function of biological age and therefore C-reactive protein measures are dependent on age. By this conceptualization, C-reactive protein measures fulfill the criteria of constituting an individualization instrument. In this case, the directed acyclic graph (c) seems to be appropriate.

Until now, we abstractly defined the attributes of an individualization instrument in the terms of conditional statistical independencies. In many clinical applications, however, statements about *metric attributes* would be useful or even necessary when the hidden trait $\boldsymbol{T}$ and the phenotype $\boldsymbol{X}$ may be understood as metric variables. In this case, we can interpret the relation between $\boldsymbol{T}$ and $\boldsymbol{X}$ in terms of measurement error models. Following our definition above, phenotypes like diseases or risk factors can be conceptualized as proxies for the underlying hidden biological traits. For example, chronological age is for sure a very good proxy of biological age, but still two individuals of age 70 may have different biological ages because of individual habits like smoking and alcohol consumption or biological traits like genetic predispositions. Using measurement error models, a metric proxy

phenotype $\boldsymbol{X}$ can be modeled using equation (1), where $\boldsymbol{T}$ is the hidden biological trait variable, and $\boldsymbol{E}_c$ is an unrelated error term regarding $\boldsymbol{T}$, following the classical measurement error model[10]

$$\boldsymbol{X} = \boldsymbol{T} + \boldsymbol{E_C} \text{ with } \boldsymbol{T} \perp\!\!\!\perp \boldsymbol{E_c}, Var(\boldsymbol{E_C}) > 0, \quad \text{classical measurement error model (CMEM)} \tag{1}$$

This is, however, not the only way to conceptualize the relation between an observable proxy and a hidden biological trait. For the example of biological age, it may be more appropriate to use equation (2) with $\boldsymbol{T}$ being the sum of $\boldsymbol{X}$ and an error term $\boldsymbol{E}$ unrelated to $\boldsymbol{X}$, also called the Berkson error model[10,11]

$$\boldsymbol{T} = \boldsymbol{X} + \boldsymbol{E_B} \text{ with } \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{E_B}, Var(\boldsymbol{E_B}) > 0, \quad \text{Berkson measurement error model (BMEM)} \tag{2}$$

Whether equation (1) or (2) is closer to the reality of the data is a conceptual decision which has to be made anew for each study design and each phenotype. Actually, the BMEM corresponds to the acyclic graph (c) whereas the CMEM is represented by the graphs (a) and (b). We can conclude directly that for an individualization instrument $\boldsymbol{Z}$ and the CMEM $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{E_C}$ holds, whereas in the BMEM $\boldsymbol{Z} \not\perp\!\!\!\perp \boldsymbol{E_B}$ is true. Hence, in the CMEM the non-informative part $\boldsymbol{E_c}$ of $\boldsymbol{X}$ regarding $\boldsymbol{T}$ is not represented in $\boldsymbol{Z}$, whereas in the BMEM the informative $\boldsymbol{E_b}$ is represented in $\boldsymbol{Z}$.

In conclusion, by understanding phenotypes as proxies for hidden biological traits, we can derive simple conditions in terms of statistical dependencies which identify variables which will be helpful in the biologically meaningful individualization of phenotypes.

3 An abstract methodology for deriving individualization metrics

After these clarifications, we will explicate the abstract methodology to derive an individualization metric in the context of a hidden trait $\boldsymbol{T}$ and a corresponding proxy phenotype $\boldsymbol{X}$ regarding using p individualization instruments $\boldsymbol{Z_1},\ldots\boldsymbol{Z_p}$. In the following paragraphs, we only discuss statistical models derived from the class of the general linear model, but extensions can be achieved easily. The goal of the methodology is to construct the individualization metric $\boldsymbol{M}$ as a linear combination of $\boldsymbol{Z_1},\ldots\boldsymbol{Z_p}$ such that the square covariance of $\boldsymbol{T}$ given $\boldsymbol{X}$ and $\boldsymbol{M}$ $Cov(\boldsymbol{M}, \boldsymbol{T}|\boldsymbol{X})^2$ is maximized or at least greater zero. Note that for the CMEM we get

$$\begin{aligned} Cov(\boldsymbol{M}, \boldsymbol{T}|\boldsymbol{X}) := Cov\left(\boldsymbol{M}, \boldsymbol{T} - \frac{Cov(\boldsymbol{T}, \boldsymbol{X})}{Var(\boldsymbol{X})}\boldsymbol{X}\right) &= Cov\left(\boldsymbol{M}, \boldsymbol{T} - \frac{Cov(\boldsymbol{T}, \boldsymbol{T} + \boldsymbol{E_C})}{Var(\boldsymbol{X})}(\boldsymbol{T} + \boldsymbol{E_C})\right) \\ &= Cov\left(\boldsymbol{M}, \boldsymbol{T} - \boldsymbol{T}\frac{Cov(\boldsymbol{T}, \boldsymbol{T})}{Var(\boldsymbol{X})}\right) + Cov\left(\boldsymbol{M}, \frac{Cov(\boldsymbol{T}, \boldsymbol{T} + \boldsymbol{E_C})}{Var(\boldsymbol{X})}\boldsymbol{E_C}\right) \\ &= Cov\left(\boldsymbol{M}, \boldsymbol{T} - \boldsymbol{T}\frac{Cov(\boldsymbol{T}, \boldsymbol{T})}{Var(\boldsymbol{X})}\right) = \left(1 - \frac{Var(\boldsymbol{T})}{Var(\boldsymbol{X})}\right)Cov(\boldsymbol{M}, \boldsymbol{T}) \end{aligned} \tag{3}$$

In the derivation, we used that $\boldsymbol{T}$ and $\boldsymbol{X}$ are linear to each other, the definition of the CMEM and that $Cov(\boldsymbol{M}, \boldsymbol{E_C}) = 0$. Analogously by using the definition of the BMEM, we get for the BMEM

$$\begin{aligned} Cov(\boldsymbol{M}, \boldsymbol{T}|\boldsymbol{X}) = Cov\left(\boldsymbol{M}, \boldsymbol{T} - \frac{Cov(\boldsymbol{T}, \boldsymbol{X})}{Var(\boldsymbol{X})}\boldsymbol{X}\right) &= Cov\left(\boldsymbol{M}, \boldsymbol{X} + \boldsymbol{E_B} - \frac{Cov(\boldsymbol{X} + \boldsymbol{E_B}, \boldsymbol{X})}{Var(\boldsymbol{X})}\boldsymbol{X}\right) \\ &= Cov\left(\boldsymbol{M}, \boldsymbol{X} + \boldsymbol{E_B} - \frac{Cov(\boldsymbol{X}, \boldsymbol{X})}{Var(\boldsymbol{X})}\boldsymbol{X}\right) = Cov(\boldsymbol{M}, \boldsymbol{E_B}) \end{aligned} \tag{4}$$

Thus, the optimization of $Cov(\boldsymbol{M}, \boldsymbol{T}|\boldsymbol{X})^2$ can be achieved by maximizing the absolute values of $Cov(\boldsymbol{M}, \boldsymbol{T})$ in CMEM and $Cov(\boldsymbol{M}, \boldsymbol{E_B})$ in the BMEM. The methodology to derive an individualization metric $\boldsymbol{M}$ and tests its possible clinical value consists of three steps.

3.1 Step 1: Predict X using $Z_1,\ldots Z_p$ and derive the prediction score Y

We use our predictors to construct a prediction score approximating $\boldsymbol{X}$, for example we predict chronological age with the $\boldsymbol{Z_i}$. As the $\boldsymbol{Z_i}$ are conditional independent of $\boldsymbol{X}$ given $\boldsymbol{T}$, any linear combination of $\boldsymbol{Z_i}$ will be conditional independent of $\boldsymbol{X}$ given $\boldsymbol{T}$ as long as the estimation of the coefficients will not introduce dependence, but this is

generally not the case in the class of linear models. Consider for example the ordinary least squares (OLS) multivariable regression of $\boldsymbol{X}$ on $\boldsymbol{Z}=(\boldsymbol{Z_1},\ldots\boldsymbol{Z_p})$ and $\boldsymbol{X}$ satisfying the CMEM. Then, the prediction score $\boldsymbol{Y}$ is given by

$$\boldsymbol{Y}=\boldsymbol{Z}\left(\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{X}\right)=\boldsymbol{Z}\left(\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T(\boldsymbol{T}+\boldsymbol{E_C})\right)=\boldsymbol{Z}\left(\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{T}\right) \tag{5}$$

as $\boldsymbol{Z}^T\boldsymbol{E_C}$ is the null-vector because of $\boldsymbol{X}\perp\!\!\!\perp\boldsymbol{Z_i}|\,\boldsymbol{T}$. Thus, $\boldsymbol{Y}$ is independent of $\boldsymbol{E_C}$. This argument is evenly true for $\boldsymbol{Y}$ being a linear combination of the principle components of $\boldsymbol{ZZ}^T$ or a linear combination of directions derived by partial least squares (PLS) algorithms or any kind of kernel regression.

In essence, every element of the linear span of $\boldsymbol{Z}$ will be itself an individualization instrument if it has a covariance with $\boldsymbol{X}$. In the consequence, the model $\boldsymbol{Y}$ can be derived via a wide range of estimation procedures including machine learning techniques like vector support machines and methods exploring the latent structure of the $\boldsymbol{Z_i}$ like PLS or principle component analyses. This argument is equally true for the BMEM. From equation (5), another important feature can be seen. The regression score $\boldsymbol{Y}$ is actually equivalent with the prediction score which one would derive when one would know the hidden trait $\boldsymbol{T}$ and regress it on the $\boldsymbol{Z_i}$. Thus, under the assumptions of OLS regression, $\boldsymbol{Y}$ is the optimal model (in terms of mean squared error) regarding $\boldsymbol{T}$ given the predictor set $\boldsymbol{Z_i}$. This attribute would be lost if any of the $\boldsymbol{Z_i}$ would violate the requirement of being conditional independence of $\boldsymbol{X}$ given $\boldsymbol{T}$. Moreover, a $\boldsymbol{Z_i}$ which would not be correlated to $\boldsymbol{X}$ would not contribute to $\boldsymbol{Y}$. Therefore, the definition of individualization instruments secures that all $\boldsymbol{Z_i}$ are correlated with $\boldsymbol{X}$.

3.2 Step 2: Regress *Y* on *X* and derive the corresponding residual variable

We derive now the prediction error orthogonalized to $\boldsymbol{X}$ by regressing the prediction score $\boldsymbol{Y}$ on $\boldsymbol{X}$ and we define this residual variable as the individualization variable M as we are most interested in information about $\boldsymbol{T}$ captured by $\boldsymbol{Y}$ given X. Therefore, we regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ and derive the prediction error in this way. We will now demonstrate for both, the CMEM and the BMEM, that

$$Cov(\boldsymbol{Y},\boldsymbol{T}|\boldsymbol{X})=Cov(\boldsymbol{T},\boldsymbol{Y}|\boldsymbol{X}) \tag{6}$$

We write down the residual variable $\boldsymbol{Y}-\hat{\boldsymbol{Y}}$ after regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$

$$\boldsymbol{M}:=\boldsymbol{Y}-\hat{\boldsymbol{Y}}=\boldsymbol{Y}-\frac{Cov(\boldsymbol{X},\boldsymbol{Y})}{Var(\boldsymbol{X})}\boldsymbol{X} \tag{7}$$

For the covariance $Cov(\boldsymbol{T},\boldsymbol{Y}|\boldsymbol{X})=Cov\left(\boldsymbol{T},\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)$ in the case of the CMEM follows:

$$\begin{aligned}Cov\left(\boldsymbol{T},\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)&=Cov(\boldsymbol{T},\boldsymbol{Y})-Cov\left(\boldsymbol{T},\hat{\boldsymbol{Y}}\right)=Cov(\boldsymbol{T},\boldsymbol{Y})-\frac{Cov(\boldsymbol{X},\boldsymbol{Y})}{Var(\boldsymbol{X})}Cov(\boldsymbol{T},\boldsymbol{X})\\&=Cov(\boldsymbol{T},\boldsymbol{Y})-\frac{Cov(\boldsymbol{T}+\boldsymbol{E_C},\boldsymbol{Y})}{Var(\boldsymbol{X})}Cov(\boldsymbol{T},\boldsymbol{T}+\boldsymbol{E_C})\\&=\left(1-\frac{Var(\boldsymbol{T})}{Var(\boldsymbol{X})}\right)Cov(\boldsymbol{T},\boldsymbol{Y})=Cov(\boldsymbol{Y},\boldsymbol{T}|\boldsymbol{X})\end{aligned} \tag{8}$$

as $\frac{Var(\boldsymbol{T})}{Var(\boldsymbol{X})}\leq 1$. For the BMEM, we get

$$\begin{aligned}Cov\left(\boldsymbol{T},\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)&=Cov(\boldsymbol{T},\boldsymbol{Y})-Cov\left(\boldsymbol{T},\hat{\boldsymbol{Y}}\right)=Cov(\boldsymbol{T},\boldsymbol{Y})-\frac{Cov(\boldsymbol{X},\boldsymbol{Y})}{Var(\boldsymbol{X})}Cov(\boldsymbol{T},\boldsymbol{X})\\&=Cov(\boldsymbol{T},\boldsymbol{Y})-\frac{Cov(\boldsymbol{T}-\boldsymbol{E_B},\boldsymbol{Y})}{Var(\boldsymbol{X})}Cov(\boldsymbol{X}+\boldsymbol{E_B},\boldsymbol{X})\\&=Cov(\boldsymbol{T},\boldsymbol{Y})-Cov(\boldsymbol{T},\boldsymbol{Y})+Cov(\boldsymbol{Y},\boldsymbol{E_B})=Cov(\boldsymbol{Y},\boldsymbol{E_B})=Cov(\boldsymbol{Y},\boldsymbol{T}|\boldsymbol{X})\neq 0\end{aligned} \tag{9}$$

Thus, the residual variable correlates with the hidden trait and contains an informational surplus in addition to the observable phenotype. Hence, the residual (7) fulfills the criteria of being an individualization metric in the context of $\boldsymbol{T}$ and $\boldsymbol{X}$. For the CMEM, the residual $\boldsymbol{M}$ is actually the optimal (in OLS terms) individualization metric reachable given the set of $\boldsymbol{Z_i}$. In the case of the BMEM, it is easy to see that $\boldsymbol{M}$ is not the optimal model, because an OLS regression of the $\boldsymbol{Z_i}$ on $\boldsymbol{T}$ directly would yield a different prediction score with higher fit to $\boldsymbol{T}|\boldsymbol{X}$. However, the

construction of the optimal model would require either observing of E_b or T. Hence, the proposed methodology cannot deliver the optimal individualization in the BMEM, but, still, the model M is informative beyond the observable phenotype and thus of potential clinical value.

In the context of individualization, the orthogonalized prediction error M is therefore not something to reduce, but to utilize as it represents the individualizing information regarding T not represented in X. Logically, a model which would predict X perfectly cannot be used for individualization and in the consequence, the model fit of a prediction rule regarding the phenotype X is not necessarily a good indicator of its meaningfulness in the sense that it delivers information not represented in X. One has to look on the informational content of the residuals after regressing Y on X.

3.3 Step 3: Demonstrate the informational content of the residual variable derived in step 2

This step is the crucial step to demonstrate the validity of the above derived metric M. Although we already have seen that the residual variable M derived in step 2 has a covariance with the hidden trait T and is therefore informative beyond the proxy X in a statistical sense, it is not clear whether the induced correlation is large enough to be practically meaningful which refers to the explained utilitarian aspect of individualization in the clinical context. Moreover, the critical prerequisite that the Z_i are individualization instruments which cannot be tested empirically may be violated, invalidating the methodology. Therefore, one needs other observable variables which correlate with T to check the validity of the derived individualization metric M. The logic in behind is simply that if M is informative for T then they should be informative for any variable related to T. The concrete procedure of step 3 is dependent on the conceptualization of the hidden trait T. For example, for biological age it makes absolutely sense to demonstrate the predictive value of the residuals derived in step 2 in survival analyses, but for Alzheimer's disease (with X being the formal diagnosis of Alzheimer's disease) it may be suitable to take neuropsychological measures or imaging markers for validation. Note that of course the parametrization of the score Y has to be numerically stable to be potentially valid in the sense of step 3. Thus, overfitting is to be avoided, e.g. by including penalty terms in the cost function.

4 Practical example: deriving a refined definition of "obesity" in the large population-based Study of Health in Pomerania (SHIP)

Now, we will apply our methodology to real-life data from the SHIP cohorts[16] and construct a refined definition of "obesity" in comparison to a pure anthropometric definition. As it is clear that obesity, while being one of most potent risk factor besides age in Western societies, is not sufficiently described by anthropometric measures,[17–19] this example is of clinical interest. The SHIP project includes population samples from north-eastern Germany with longitudinal and comprehensive medical phenotyping. Here, we will use the SHIP-0 cohort ($n = 4308$) which was sampled between 1997 and 2001 and had follow-up survival data until 2015. The investigations were performed in accordance with the Declaration of Helsinki, including written informed consent of all participants. The survey and study methods of the SHIP studies were approved by the institutional review boards of the University of Greifswald. For details on SHIP, see Völzke et al.[16] and the Supplementary material, describing sampling strategies, measurements and phenotyping. The sample characteristics of SHIP-0

Table 1. Descriptive statistics for the utilized SHIP-0 cohort.

	Women ($n = 1826$)	Men ($n = 1710$)
Age, mean(SD)	48.47 (16.22)	50.66 (16.47)
Waist circumference in cm, mean(SD)	83.19 (12.89)	95.84 (11.51)
Body mass index in kg/m^2, mean(SD)	26.97 (5.28)	27.70 (3.97)
Triglycerides in mmol/L, mean(SD)	1.55 (0.96)	2.08 (1.39)
Systolic Blood pressure in mmHg, mean(SD)	129.55 (20.84)	142.35 (19.23)
Glycated hemoglobin in percentage, mean(SD)	5.32 (0.84)	5.48 (0.88)
Diabetes (%)	7.78	10.54
Smoking (%)	27.29	34.49

SD: standard deviation.

can be found in Table 1. All data used here can be accessed via the data application procedure on www.community-medicine.de free of charge.

4.1 Step 1: Predict the waist circumference using measures indicative of metabolic health and derive the prediction score *Y*

For the goal of deriving a refined measure of obesity, we extracted eight measures associated with metabolic health. These measures were the glycated hemoglobin percentage, systolic blood pressure, cystatin C, C-reactive protein measures (high-sensitive), triglycerides, total cholesterol to high density lipoprotein cholesterol ratio, red blood cell counts and white blood cell counts. Complete information on these variables was available in 3547 cases (men = 1716, women = 1831). We conceptualize the waist circumference as the phenotype $\boldsymbol{X}$, the obesity-related metabolic disruptions as $\boldsymbol{T}$ and the named predictors as $\boldsymbol{Z_i}$. The $\boldsymbol{Z_i}$ are plausible individualization instruments, because they are influenced by the metabolic disruptions caused by obesity. We assume that there is no path directly from the waist circumference to any $\boldsymbol{Z_i}$, assuming thus conditional independence of the $\boldsymbol{Z_i}$ and the waist circumference given $\boldsymbol{T}$. In essence, this means we conceptualize the variables in the sense of the BMEM where the hidden trait $\boldsymbol{T}$ is influenced not only by obesity measured by the waist circumference, but also by other traits like genetic variants or physical activity.

In the next step, a multivariable regression was fitted with the waist circumference as outcome variable and the named variables as predictors. We used the multivariable fractional polynomial (MFP) approach[20] to model potential non-linearity of these variables and fitted the model for men and women separately, as it is plausible that the used variables may not behave similarly for men and women.[21] We allowed the transformations X^{-2}, X^{-1}, $X^{-0.5}$, log(X), $X^{0.5}$ X, X^2, and X^3 with maximal five cycles of iterations. The model reached convergence after three cycles of iterations in both sexes. For women, the model reached an adjusted R-squared of 0.43 and for men 0.33. Ten-fold internal cross-validation using 15 repetitions with newly randomized separations supported the model fit and showed no indication for overfitting as one could expect with around 200 observations per predictor. The prediction scores were different for men and women in their parametrization and for both sexes, the final model included several non-linear transformations. The full parametrization can be found in the Supplementary material (Tables S1 and S2). Note that the mediocre R-squared is nothing that concerns us here as it is not the crucial criterion for a good model in our sense as explained above.

The chosen variables are standard measures of health and widely available over cohort studies, so replication can be easily conducted. Of course, there may be other variables enhancing the informational content of the corresponding residual variables, but our aim here is to show that even with standard health indicators a refined definition of obesity is possible, demonstrating on the way the usefulness of our approach to individualization.

4.2 Step 2: Regress the prediction score on the waist circumference and chronological age and derive the corresponding residual variable *M*

In the next step, we regressed the prediction score described above on the waist circumference and additionally on the chronological age, separately for men and women in an ordinary least squares regression. Chronological age was used as covariate here to derive a score independent of chronological age. Subsequently, we calculated the corresponding residual variables $\boldsymbol{M}$. Conceptually, these residuals describe the difference between the individual waist circumference and the prediction score which was expected given his/her true waist circumference and her/his chronological age. Thus, these residuals are differentiating between two persons of the same waist circumference and the residual variable can be seen as individualization metric. Our theoretical arguments predict that these differentiation correlates with the true difference in metabolic health between person of the same anthropomorphic obesity measure. This claim which is based on not easily falsifiable assumptions, of course, has to be tested on validity which is done in the third step.

4.3 Step 3: Demonstrate the informational content of the residual prediction score on prospective survival data

Now, we show that $\boldsymbol{M}$ is indeed informative regarding survival (for details on the sampling of the survival data and the definition of cardiovascular mortality, see the ''Extended Methods section'' in the Supplementary Material; for further results see Supplementary Tables S3 and Supplementary Figures S1 and S2). From the individuals with complete covariate vector, a total of 659 individuals died in the follow-up interval and the analyzed failure event

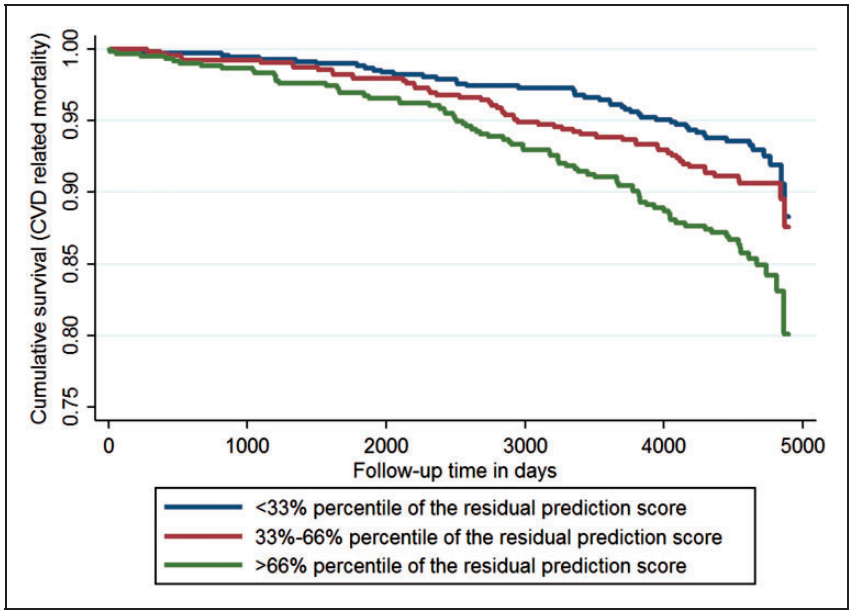

Figure 1. Kaplan–Meier curves regarding cardiovascular related mortality of the tertiles of the residual after regressing prediction score on the chronological age and the WC. The curves are statistically different (log-rang test: $\chi^2(2) = 8.13$, $p = 0.017$).

was cardiovascular related mortality (206 deaths). In Cox regressions, the residual variable was predictive for survival with a hazard ratio per point of 1.050 (95%-confidence interval: (1.023–1.077), $p = 0.0002$), adjusted for sex and age. For the visualization of the effect, see Kaplan–Meier curves (Figure 1) displaying the sex specific tertiles of the residual scores. In addition, one can find in the Supplementary Material, Kaplan–Meier curves for the age-groups above 60 years and below 60 years (Figures S1 and S2). Clearly, the discriminative power of the individualization metric is in the older individuals whereas in the younger individuals there was no indication of a predictive value for the categorized residual variable. However, the low number of deaths in the younger age-group means that this can be also merely a problem of statistical power. To achieve easier interpretation of the individualization metric $\boldsymbol{M}$, we rescaled $\boldsymbol{M}$ such that one unit represents the same increase in risk as one unit on the usual BMI scale (kg/m^2) conditional on age and sex. Now, we just summed up the residual variable and the standard BMI variable for a measure of metabolic health that we call the *metabolic BMI*. We can now refine common definitions of obesity, for example by applying a cut-off at 35. For a graphical representation, see Figure 2. This procedure ends in a canonical definition of 'healthy obesity' and an 'unhealthy lean' status (see Figure 2). Indeed, when testing this classification in prospective survival analyses, the 'healthy' obese individual had no higher risk (HR: 0.99, 95% CI: 0.37–2.70, $p = 0.992$) than "healthy" subjects with a BMI < 35, while the class of "unhealthy lean" individuals ($n = 205$) with a BMI < 35 had a HR of 2.19 (95% CI: 1.37–3.47, $p < 0.001$). Naturally, the group with a BMI > 35 and a metabolic BMI > 35 had the highest hazard ratio of 3.76 (95% CI: 2.34–5.98, $p < 0.001$). This mirrors the fact that the BMI itself is predictive for cardiovascular mortality. As a limitation, however, it should be noted that the "healthy obesity" group was rather small ($n = 65$). Thus, the reported results as indicated by the wide confidence intervals have to be treated with care.

In conclusion, by applying our methodology we derived a refined definition of obesity which was statistically superior in the prediction of cardiovascular death. Thus, regarding the risk of dying from cardiovascular causes, our metric was able to differentiate individuals showing the same anthropometry and fulfill thereby our definition of individualization.

5 Consequences for study design and statistical analyses

The basic message from the explicated methodology above is that multivariate prediction scores contain more information about hidden traits than usually utilized and that this information is extractable and can be used for individualization. This is in its core good news: it is already possible for researchers to go beyond the usual classification systems of health and disease, delivering individualized metrics to the clinical sciences. However,

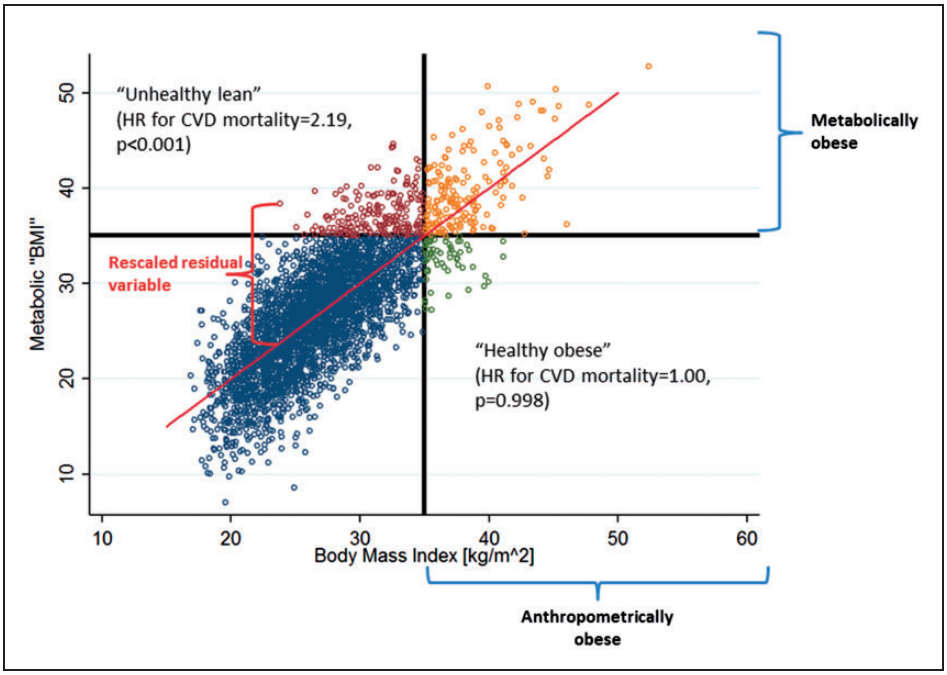

Figure 2. Graphical representation of the refined definition of obesity. The metabolic BMI is the sum of the rescaled residual variable resulting from regressing the prediction score on age and the anthropometric WC such as one point increase in this residual scale equals one point increase in BMI regarding the hazard of dying from cardiovascular causes in the SHIP sample. The red line indicates the identity.

doing so requires a paradigm shift in the way we evaluate and we build prediction scores. Until now, researchers are mainly interested in good and stable model fit regarding the phenotypes under consideration.[22–24] In contrast, we argue that the informative "individualizing" content of a score lays within its prediction error (orthogonalized to the modeled phenotype) which can be tested on a set of validation variables (step 3). From a more conceptual viewpoint, one would not only ask how many cases and controls were correctly assigned to their classes, but what information is given by the cases that were wrongly classified and what information is given by the distance to the hyperplane which separates the groups. Our arguments above show that this could be actually very informative. For a clinical example, it may be that in the case of cancer, misclassification is in line with a good prognosis. Of course, such bold claims must be backed up by the corresponding data. This would be only possible if such variables are part of the study design. Thus, it is clear that the procedure of step 3 should be considered in the study design (ensuring that the necessary information is acquired). Moreover, the validation procedure should be strictly pre-specified allowing falsification and a clear definition of what an individualized measure should satisfy in the context of the research question and clinical application.

The construction of individualized phenotypes thus is a further example of the principle "no biology in, no biology out". Given the fact that most of our phenotypes cannot be identified with the biology in behind, the goal of individualization is not reachable by data-driven procedures alone. The construction of individualized phenotypes implies a *conceptual clarification* of the relation among the observed variables and a conceptualization of the implicitly modeled hidden trait.

Our arguments have also an impact on the way the predictors are selected in the modeling process of the phenotype. Normally, one would choose reliable predictors which correlate strongly with the observable phenotype.[25–27] We think that this procedure is likely to lead to a set of predictors not consisting exclusively of individualization instruments. To make this point clearer, we will introduce here the term "*conceptual overfitting*" (see Figure 3). Conceptual overfitting arises when in constructing the model $\boldsymbol{Y}$ a predictor $\boldsymbol{Z_i}$ is included which is related to the phenotype, but is not independent of the observable variable $\boldsymbol{X}$ conditional on the hidden trait $\boldsymbol{T}$. In this case, the model fit for $\boldsymbol{Y}$ regarding $\boldsymbol{X}$ would be truly higher if $\boldsymbol{Z_i}$ is included, but it would not be necessarily beneficial for the covariance of the residual variable with the hidden trait $\boldsymbol{T}$. For an example, consider the risk factor BMI which can be seen as a proxy of metabolic health. However, the BMI is biased regarding the true

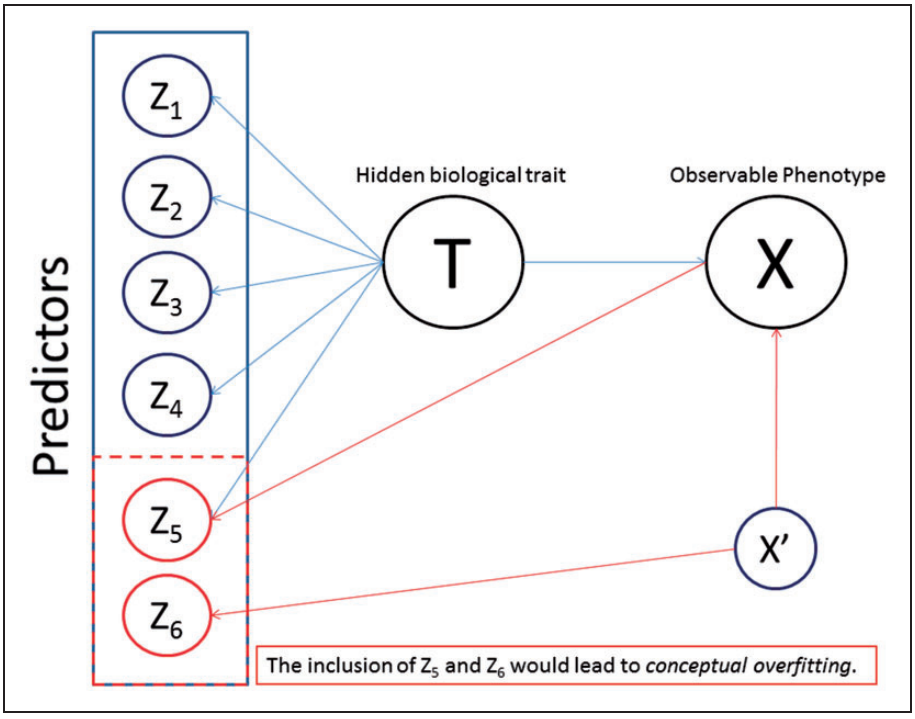

Figure 3. Graphical representation of *conceptual overfitting*: Z_5 and Z_6 would increase the fit of a prediction model regarding X, but would invalidate the conditional independence assumption on which the informative content of the residual variable relies.

metabolic status by muscle mass.[19] Predicting the BMI via metabolomics and including metabolites indicative for muscle mass (creatinine, branched chain amino acids) would lead to a higher R-squared and a truly better model for the BMI, but would result in an equally biased measurement for metabolic health. Excluding the metabolites related to muscle mass may therefore lead to lower fit, but to higher informational content regarding the metabolic health of an individual. Hence, predictor selection should be done if possible on theoretical grounds. As the conditional independence assumptions given for the definition of individualization instruments can be seen in the context of causal[15] inference theory, one could apply directed acyclic graphs to choose an appropriate set of predictors. Of course, it is often difficult and in the case of big data modeling often impossible to do so comprehensively. Still, we believe it is important to notice that predictor selecting maximizing model fit is not a sensible procedure when it comes to the individualization of risk factors and phenotypes.

6 Conclusions

We delivered a definition of individualization and then transferred it to statistical terms by defining individualization in the context of a phenotype not perfectly correlated with the underlying biology. Thus, it is clear that individualization is context-dependent. The context of every individualization (the proxy-phenotype, the hidden biological trait) has to be always explicated and the respective choices have to be motivated, otherwise the term "individualization" is without proper meaning. On the ground of these definitions, we proposed an abstract methodology compatible with a wide range of estimation procedures, targeting the individualization of phenotypes. We utilized measurement error theory to demonstrate that the prediction error orthogonalized to the predicted phenotype is always correlated to the underlying, hidden trait if the predictors fulfill the criteria of being individualization instruments. The central prerequisite states that the predictors must be conditionally independent on the phenotype given the hidden trait variable which implies that the conceptual relation of the predictors to the modeled phenotype has to be considered when applying our workflow. As the underlying theoretical assumptions will be often hard to test empirically, it is important to pre-specify testable conditions under which the newly derived phenotype is considered to be valid (see step 3). We utilized the methodology with success on epidemiological data from the SHIP cohort to construct a meaningful refined definition of obesity. In conclusion, when individualization is the goal of statistical

modeling, the prediction error is not something to avoid but to utilize, leading to the meaningful differentiation of individuals showing the same observable phenotype. We hope that our arguments are one step forward on the way of individualized phenotypes into everyday clinics, a critical prerequisite for the facilitation of individualized medicine.[28]

Authors' contribution

JH wrote the manuscript and developed the methodology. SF checked all mathematical derivations. HJG, SF, JK, KW, GF, BH, and TK took part in the development of the concepts. HV, TK, MN, and NF contributed crucially to the design and the phenotyping of the SHIP studies. All authors reviewed the final manuscript.

Acknowledgements

The contribution to data collection performed by study nurses, study physicians, interviewers, and laboratory workers is gratefully acknowledged. We are also appreciative of the important support of IT and computer scientists, health information managers and administration staff. We also thank all study participants whose personal dedication and commitment made this project possible.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was part of the GANI_MED (Greifswald Approach to Individualized Medicine) project, which is funded by the Federal Ministry of Education and Research [grant number 03IS2061A]. SHIP is part of the Community Medicine Research Network of the University Medicine Greifswald, Germany, which is supported by the German Federal State of Mecklenburg-West Pomerania. GF is supported by the European Union's Horizon 2020 research and innovation program (grant number 633589).

ORCID iD

Johannes Hertel http://orcid.org/0000-0002-7641-0132

Supplemental material

Supplemental material for this article is available online.

References

1. Tavakolpour S. Towards personalized medicine for patients with autoimmune diseases: opportunities and challenges. *Immunol Lett* 2017; (in press). DOI: 10.1016/j.imlet.2017.08.002
2. O'Donnell JC. Personalized medicine and the role of health economics and outcomes research: issues, applications, emerging trends, and future research. *Value Health* 2013; **16**: S1–S3.
3. Schleidgen S, Klingler C, Bertram T, et al. What is personalized medicine: sharpening a vague term based on a systematic literature review. *BMC Med Ethics* 2013; **14**: 55.
4. Jylhävä J, Pedersen NL and Hägg S. Biological age predictors. *EBioMedicine* 2017; **21**: 29–36.
5. Kramer CK, Zinman B and Retnakaran R. Are metabolically healthy overweight and obesity benign conditions? A systematic review and meta-analysis. *Ann Intern Med* 2013; **159**: 758–769.
6. Blüher S and Schwarz P. Metabolically healthy obesity from childhood to adulthood – does weight status alone matter? *Metab Clin Exp* 2014; **63**: 1084–1092.
7. Antoñanzas F, Juárez-Castelló CA and Rodríguez-Ibeas R. Some economics on personalized and predictive medicine. *Eur J Health Econ* 2015; **16**: 985–994.

8. Hertel J, Friedrich N, Wittfeld K, et al. Measuring biological age via metabonomics: the metabolic age score. *J Proteome Res* 2016; **15**: 400–410.
9. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013; **14**: R115.
10. Buonaccorsi JP. *Measurement error: models, methods, and applications*. Boca Raton, FL: CRC Press, 2010.
11. Berkson J. Are there two regressions? *J Am Stat Assoc* 1950; **45**: 164–180.
12. Perna L, Zhang Y, Mons U, et al. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin Epigenetics* 2016; **8**: 64.
13. Habes M, Janowitz D, Erus G, et al. Advanced brain aging: relationship with epidemiologic and genetic risk factors, and overlap with Alzheimer disease atrophy patterns. *Transl Psychiatr* 2016; **6**: e775.
14. Menni C, Kiddle SJ, Mangino M, et al. Circulating proteomic signatures of chronological age. *J Gerontol A Biol Sci Med Sci* 2015; **70**: 809–816.
15. Pearl J. Causal inference in statistics: an overview. *Stat Survey* 2009; **3**: 96–146.
16. Völzke H, Alte D, Schmidt CO, et al. Cohort profile: the study of health in Pomerania. *Int J Epidemiol* 2011; **40**: 294–307.
17. Global Burden of Metabolic Risk Factors for Chronic Diseases Collaboration (BMI Mediated Effects), Lu Y, Hajifathalian K, et al. Metabolic mediators of the effects of body-mass index, overweight, and obesity on coronary heart disease and stroke: a pooled analysis of 97 prospective cohorts with 1.8 million participants. *Lancet* 2014; **383**: 970–983.
18. Ahima RS and Lazar MA. Physiology. The health risk of obesity – better metrics imperative. *Science* 2013; **341**: 856–858.
19. Schneider HJ, Friedrich N, Klotsche J, et al. The predictive value of different measures of obesity for incident cardiovascular events and mortality. *J Clin Endocrinol Metab* 2010; **95**: 1777–1785.
20. Royston P and Sauerbrei W. *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Chichester, UK; Hoboken, NJ: John Wiley, 2008.
21. Palmer BF and Clegg DJ. The sexual dimorphism of obesity. *Mol Cell Endocrinol* 2015; **402**: 113–119.
22. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. New York, NY: Springer, 2009.
23. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**: 128–138.
24. Roy J, Shou H, Xie D, et al. Statistical methods for cohort studies of CKD: prediction modeling. *Clin J Am Soc Nephrol* 2017; **12**: 1010–1017.
25. Wiegand RE. Performance of using multiple stepwise algorithms for variable selection. *Stat Med* 2010; **29**: 1647–1659.
26. Sauerbrei W, Royston P and Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med* 2007; **26**: 5512–5528.
27. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer, 2001.
28. Grabe HJ, Assel H, Bahls T, et al. Cohort profile: Greifswald approach to individualized medicine (GANI_MED). *J Transl Med* 2014; **12**: 144.