

Health outcome prediction using multiple perturbations

Wen-Chung Lee, MD, PhD*

Abstract

Public health workers and medical practitioners are frequently required to make predictions regarding various health outcomes. However, a prediction with nearly 100% certainty is seldom possible.

If a person has a health outcome of concern or is in the process of developing the outcome, many attributes of that person may undergo subtle changes—the perturbations. We propose a method, namely “prediction using multiple perturbations” and investigate its asymptotic properties when the number of attributes tends to infinity. This is a proof-of-concept study.

The proposed method can predict the health outcome of a person to near certainty if personal data with billions or trillions of attributes can be collected and 4 conditions (described subsequently in this paper) are met.

Collecting personal data with billions or trillions of attributes may someday become possible in the current era of big data. If such information can be obtained, theoretically we can predict the health outcome of a person to near certainty.

Abbreviation: PUMP = prediction using multiple perturbations.

Keywords: big data, biostatistics, data mining, outcome prediction, predictive analytics

1. Introduction

Public health workers and medical practitioners are frequently required to make predictions regarding various health outcomes. For example, they may be required to predict whether a 65-year-old woman with a 50-year history of alcohol drinking, betel nut chewing, and cigarette smoking will develop oral cancer in 1 year.

Editor: Jimmy T. Efrid.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The dataset supporting the conclusions of this article is included within the article.

Funding: This paper is supported by grants from the Ministry of Science and Technology in Taiwan (MOST 104-2314-B-002-118-MY3, MOST 105-2314-B-002-049-MY3, MOST 108-3017-F-002-001, MOST 108-2314-B-002-127-MY3), and the Innovation and Policy Center for Population Health and Sustainable Environment (Population Health Research Center, PHRC) from Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan (NTU-109L9003008). No additional external funding was received for this study. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The author declares that he has no competing interests.

Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan.

* Correspondence: Wen-Chung Lee, Rm. 536, No. 17, Xuzhou Rd., Taipei 100, Taiwan (e-mail: wenchung@ntu.edu.tw).

Copyright © 2020 the Author(s). Published by Wolters Kluwer Health, Inc. This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial License 4.0 (CCBY-NC), where it is permissible to download, share, remix, transform, and buildup the work provided it is properly cited. The work cannot be used commercially without permission from the journal.

How to cite this article: Lee WC. Health outcome prediction using multiple perturbations. *Medicine* 2020;99:2(e18664).

Received: 19 April 2019 / Received in final form: 27 October 2019 / Accepted: 4 December 2019

<http://dx.doi.org/10.1097/MD.00000000000018664>

They may also be required to predict whether a young man without previous medical history will survive an emergency operation for a ruptured dissecting aortic aneurysm. Many risk or prognostic factors have been found for nearly every health outcome. On the basis of such factors, a risk prediction model (e.g., the Framingham score for cardiovascular risk^[1] or a disease-staging system [e.g., the International Federation of Gynecology and Obstetrics staging for cervical cancer prognosis])^[2] can be constructed to make accurate predictions. However, a prediction with approximately 100% certainty is seldom possible.

We report a new avenue for health outcome prediction. The method hinges on collecting multiple “perturbations”^[3,4] of the health outcome of interest. Notably, if a person has the health outcome of concern (e.g., clinically diagnosed liver cancer) or is in the process of developing the outcome (e.g., a small malignant liver tumor not yet manifested clinically), many attributes of that person may undergo subtle changes (referred to as the “perturbed” attributes in this paper). For example, the person’s physical/emotional characteristics, physiological/biochemical profiles, various “omics” (e.g., epigenomics, transcriptomics, proteomics, metabolomics, and exposomics), behavior patterns, social activities, and data pertaining to that person. Notably, changes induced by the health outcome may be nondeterministic (i.e., only vary in probability), and the magnitudes of the changes may be minuscule. We utilize such health big data and propose a method, namely “prediction using multiple perturbations” (PUMP). This is a proof-of-concept study. In this paper, we investigate the asymptotic properties of PUMP when the number of attributes tends to infinity.

2. Methods

2.1. Training samples

To train a PUMP to predict a health outcome (Y), we need a case sample (people with Y; with a sample size of n_1 indexed by j) and

2 independent control samples (people without Y; with a sample size of $n_{0,1}$ indexed by k_1 , and a sample size of $n_{0,2}$ indexed by k_2 , respectively). Information on a total of m attributes (indexed by i) is accrued, for all people in the case sample ($X_{i,j}^{\text{case}}$) and the 2 control samples ($X_{i,k_1}^{\text{control-1}}$ and $X_{i,k_2}^{\text{control-2}}$). The attributes are assumed to be binary (0 or 1); otherwise, we censor values that are excessively high to an (arbitrarily defined) high limit and those that are excessively too low to a low limit, followed by mapping all values to the unit interval. The point of doing such censoring is to bound the ranges of all attributes to between zero and one: $0 \leq X_{i,j}^{\text{case}}, X_{i,k_1}^{\text{control-1}}, X_{i,k_2}^{\text{control-2}} \leq 1$, for all i, j, k_1 , and k_2 . Next, we calculate the mean attributes for the 3 training samples:

$$\bar{X}_i^{\text{case}} = \frac{1}{n_1} \times \sum_{j=1}^{n_1} X_{i,j}^{\text{case}}, \quad \bar{X}_i^{\text{control-1}} = \frac{1}{n_{0,1}} \times \sum_{k_1=1}^{n_{0,1}} X_{i,k_1}^{\text{control-1}}, \quad \text{and}$$

$$\bar{X}_i^{\text{control-2}} = \frac{1}{n_{0,2}} \times \sum_{k_2=1}^{n_{0,2}} X_{i,k_2}^{\text{control-2}}, \quad \text{respectively, for } i=1, 2, \dots, m.$$

2.2. The prediction method

To make a prediction for a new person, we accrue the information of the corresponding attributes pertaining to the person: X_i^{new} , for $i=1, 2, \dots, m$. We calculate a perturbation score at each and every attribute for the new person: $S_i^{\text{new}} = (X_i^{\text{new}} - \bar{X}_i^{\text{control-1}}) \times (\bar{X}_i^{\text{case}} - \bar{X}_i^{\text{control-2}})$, for $i=1, 2, \dots, m$. These scores are then averaged across the attributes to yield $\bar{S}^{\text{new}} = \frac{1}{m} \times \sum_{i=1}^m S_i^{\text{new}}$. We predict the new person to eventually develop Y if his/her average score is larger than a certain threshold value (a small positive number near zero, to be discussed later); otherwise, we predict the person to not develop Y.

2.3. Ethical review

This paper is a methodological study and does not involve the enrollment of study subjects. Ethical approval is not necessary.

3. Results

As the number of attributes tends to infinity ($m \rightarrow \infty$), the probability of a correct prediction using the proposed method tends to 1 if the following 4 conditions are met:

- (C1) An attribute may or may not be perturbed by the process of developing the health outcome. For a perturbed attribute, the mean for the people with the outcome and the mean for the people without the outcome flank the mean for those in the process of developing the outcome. For a non-perturbed attribute, the means are identical for all 3 types of people.
- (C2) The number of perturbed attributes has an asymptotic growth rate no lower than that of the total number of attributes.
- (C3) The mean of a perturbed attribute for the people in the process of developing the outcome is bounded away from the mean for those without the outcome.
- (C4) An attribute correlates with a bounded number of other attributes, for the people with the outcome, the people in the process of developing the outcome, or the people without the outcome.

Note that (C1), (C2), and (C3) above concern the relation between the many attributes to be collected and the health

outcome under concern, whereas (C4) above concerns the relation between the attributes themselves.

3.1. A proof that the probability of a correct prediction tends to one

Let the mean (variance) of the i th attribute be denoted by $\mu_i^{\text{with Y}}$ ($v_i^{\text{with Y}}$) for the people with Y, $\mu_i^{\text{developing Y}}$ ($v_i^{\text{developing Y}}$) for the people in the process of developing Y, and $\mu_i^{\text{without Y}}$ ($v_i^{\text{without Y}}$) for the people without Y, respectively. Because all attributes are between zero and one inclusive, we have that $0 \leq \mu_i^{\text{with Y}}, \mu_i^{\text{developing Y}}, \mu_i^{\text{without Y}} \leq 1$, respectively, for all i . For attributes that are bounded between zero and one, the variances are the largest when they are Bernoulli distributed (either 0 or 1, but not between). Therefore, we have that $v_i^{\text{with Y}} \leq \mu_i^{\text{with Y}} \times (1 - \mu_i^{\text{with Y}})$, $v_i^{\text{developing Y}} \leq \mu_i^{\text{developing Y}} \times (1 - \mu_i^{\text{developing Y}})$, and $v_i^{\text{without Y}} \leq \mu_i^{\text{without Y}} \times (1 - \mu_i^{\text{without Y}})$, respectively, for all i .

From C1, let $I_i = 1$ indicate that the i th attribute is a perturbed attribute ($\mu_i^{\text{with Y}} > \mu_i^{\text{developing Y}} > \mu_i^{\text{without Y}}$ or $\mu_i^{\text{with Y}} < \mu_i^{\text{developing Y}} < \mu_i^{\text{without Y}}$), and $I_i = 0$ otherwise ($\mu_i^{\text{with Y}} = \mu_i^{\text{developing Y}} = \mu_i^{\text{without Y}}$). The number of perturbed attributes is $\psi_m = \sum_{i=1}^m I_i$, which is a function of m . From C2, there exist positive constants π and M such that $\psi_m \geq \pi \times m > 0$ for all $m \geq M$. From C3, there exists a positive constant ξ such that $|\mu_i^{\text{developing Y}} - \mu_i^{\text{without Y}}| \geq \xi$ for all i with $I_i = 1$.

Because of the independence between $(X_i^{\text{new}} - \bar{X}_i^{\text{control-1}})$ and $(\bar{X}_i^{\text{case}} - \bar{X}_i^{\text{control-2}})$ (the former and the latter being based on different people), the means and variances of the perturbation score at the i th attribute for a new person are $E(S_i^{\text{new}}) = E(X_i^{\text{new}} - \bar{X}_i^{\text{control-1}}) \times E(\bar{X}_i^{\text{case}} - \bar{X}_i^{\text{control-2}}) = (\mu_i^{\text{new}} - \mu_i^{\text{without Y}}) \times (\mu_i^{\text{with Y}} - \mu_i^{\text{without Y}})$ and

$$\begin{aligned} \text{Var}(S_i^{\text{new}}) &= \text{Var}(X_i^{\text{new}} - \bar{X}_i^{\text{control-1}}) \times E^2(\bar{X}_i^{\text{case}} - \bar{X}_i^{\text{control-2}}) \\ &\quad + E^2(X_i^{\text{new}} - \bar{X}_i^{\text{control-1}}) \times \text{Var}(\bar{X}_i^{\text{case}} - \bar{X}_i^{\text{control-2}}) \\ &\quad + \text{Var}(X_i^{\text{new}} - \bar{X}_i^{\text{control-1}}) \times \text{Var}(\bar{X}_i^{\text{case}} - \bar{X}_i^{\text{control-2}}) \\ &\leq V_{\mu,n} \end{aligned}$$

respectively, where

$$\begin{aligned} V_{\mu,n} &= \left[\mu_i^{\text{new}} \times (1 - \mu_i^{\text{new}}) + \frac{\mu_i^{\text{without Y}} \times (1 - \mu_i^{\text{without Y}})}{n_{0,1}} \right] \\ &\quad \times (\mu_i^{\text{with Y}} - \mu_i^{\text{without Y}})^2 + (\mu_i^{\text{new}} - \mu_i^{\text{without Y}})^2 \\ &\quad \times \left[\frac{\mu_i^{\text{with Y}} \times (1 - \mu_i^{\text{with Y}})}{n_1} + \frac{\mu_i^{\text{without Y}} \times (1 - \mu_i^{\text{without Y}})}{n_{0,2}} \right] \\ &\quad + \left[\mu_i^{\text{new}} \times (1 - \mu_i^{\text{new}}) + \frac{\mu_i^{\text{without Y}} \times (1 - \mu_i^{\text{without Y}})}{n_{0,1}} \right] \\ &\quad \times \left[\frac{\mu_i^{\text{with Y}} \times (1 - \mu_i^{\text{with Y}})}{n_1} + \frac{\mu_i^{\text{without Y}} \times (1 - \mu_i^{\text{without Y}})}{n_{0,2}} \right] \end{aligned}$$

is a function of the means of the attributes and the sample sizes of the training data.

From C4, assume that any attribute can correlate with at most C other attributes, for either the people with Y , the people in the process of developing Y , or the people without Y . Then, we have that $\sum_{i \neq j} \text{Cov}(S_i^{\text{new}}, S_j^{\text{new}}) \leq C \times V_{\mu,n}$, for all i .

We now calculate the mean of the average score for the new person as a function of m . We have that

$$\begin{aligned} E(\bar{S}_m^{\text{without } Y}) &= \frac{1}{m} \times \sum_{i=1}^m E(S_i^{\text{without } Y}) \\ &= \frac{1}{m} \times \sum_{i=1}^m (\mu_i^{\text{without } Y} - \mu_i^{\text{without } Y}) \\ &\quad \times (\mu_i^{\text{with } Y} - \mu_i^{\text{without } Y}) \\ &= 0 \end{aligned}$$

for all m if the new person is without Y , and that

$$\begin{aligned} E(\bar{S}_m^{\text{developing } Y}) &= \frac{1}{m} \times \sum_{i=1}^m E(S_i^{\text{developing } Y}) \\ &= \frac{1}{m} \times \sum_{i=1}^m (\mu_i^{\text{developing } Y} - \mu_i^{\text{without } Y}) \\ &\quad \times (\mu_i^{\text{with } Y} - \mu_i^{\text{without } Y}) \\ &\geq \frac{1}{m} \times \sum_{i=1}^m I_i \times (\mu_i^{\text{developing } Y} - \mu_i^{\text{without } Y})^2 \\ &\geq \frac{\psi_m \times \xi^2}{m} \\ &\geq \pi \times \xi^2 \end{aligned}$$

for all $m \geq M$ if the new person is developing Y . As for the variance of the average score for the new person, irrespectively of whether he/she is without Y or developing Y we have that

$$\begin{aligned} \text{Var}(\bar{S}_m^{\text{new}}) &= \frac{1}{m^2} \times \sum_{i=1}^m \left[\text{Var}(S_i^{\text{new}}) + \sum_{i' \neq i} \text{Cov}(S_i^{\text{new}}, S_{i'}^{\text{new}}) \right] \\ &\leq \frac{1+C}{m} \times V_{\mu,n} \end{aligned}$$

for all m .

For any t in $(0, \pi \times \xi^2)$ and any $m \geq M$, the probability of a correct prediction for a person without Y is $Pr(\bar{S}_m^{\text{without } Y} \leq t) \geq 1 - \frac{1}{t^2} \times \text{Var}(\bar{S}_m^{\text{without } Y}) \geq 1 - \frac{1+C}{m \times t^2} \times V_{\mu,n}$ and the probability of a correct prediction for a person developing Y is $Pr(\bar{S}_m^{\text{developing } Y} > t) \geq 1 - \frac{1}{(t - \pi \times \xi^2)^2} \times \text{Var}(\bar{S}_m^{\text{developing } Y}) \geq 1 - \frac{1+C}{m \times (t - \pi \times \xi^2)^2} \times V_{\mu,n}$.

A simple numerical analysis shows that $V_{\mu,n} \leq 0.3462$. Therefore, we see that as the number of attributes tends to infinity ($m \rightarrow \infty$), the probability of a correct prediction tends to one.

3.2. Probability of a correct prediction and number of attributes needed

Assuming $\mu_i^{\text{with } Y} \approx \mu_i^{\text{new}} \approx \mu_i^{\text{without } Y}$, we have that $V_{\mu,n} \approx \frac{1}{16} \times \left(1 + \frac{1}{n_{0,1}}\right) \times \left(\frac{1}{n_1} + \frac{1}{n_{0,2}}\right)$. Let $t = \frac{1}{2} \times \pi \times \xi^2$, the probability of a correct prediction is then $Pr(\bar{S}_m^{\text{without } Y} \leq t) = Pr(\bar{S}_m^{\text{developing } Y} > t) \geq 1 - \frac{1+C}{4 \times m \times \pi^2 \times \xi^4} \times \left(1 + \frac{1}{n_{0,1}}\right) \times \left(\frac{1}{n_1} + \frac{1}{n_{0,2}}\right)$

(assuming $m \geq M$). Figure 1 shows that as the number of attributes increases, the lower bound for the probability of a correct prediction increases.

To control a false positive rate (the probability of a wrong prediction for a person without the outcome) no larger than α ($0 < \alpha < 1$) and a false negative rate (the probability of a wrong prediction for a person developing the outcome) no larger than β ($0 < \beta < 1$), we can set the threshold value at $t^* = \arg \min_{t \text{ in } (0, \pi \times \xi^2)} \left\{ \max \left[\frac{1}{\alpha \times t^2}, \frac{1}{\beta \times (t - \pi \times \xi^2)^2} \right] \right\} = \frac{\sqrt{\beta}}{(\sqrt{\alpha} + \sqrt{\beta})} \times \pi \times \xi^2$.

The total number of attributes needed is then $m_{[\alpha, \beta]} \approx \left(\frac{1}{\sqrt{\alpha}} + \frac{1}{\sqrt{\beta}}\right)^2 \times \frac{1+C}{16 \times \pi^2 \times \xi^4} \times \left(1 + \frac{1}{n_{0,1}}\right) \times \left(\frac{1}{n_1} + \frac{1}{n_{0,2}}\right)$ (assuming $m_{[\alpha, \beta]} \geq M$). Table 1 shows that an extremely large number of attributes ($m_{[0.01, 0.01]}$) is required to control the false positive and the false negative rates both no larger than 0.01. In Fig. 1, the lower bound for the probability of a correct prediction is larger than zero for $m > m_{[1,1]} \approx \frac{1+C}{4 \times \pi^2 \times \xi^4} \times \left(1 + \frac{1}{n_{0,1}}\right) \times \left(\frac{1}{n_1} + \frac{1}{n_{0,2}}\right)$ (assuming $m_{[1,1]} \geq M$).

4. Discussion

Conventional asymptotic analysis assumes the number of subjects to tend to infinity. Hall et al^[5] and Ahn et al^[6] proposed an alternative approach assuming the number of “dimensions” (corresponding to “attributes” in this paper), instead, to tend to infinity. Previously, we built on this alternative asymptotic to develop new methods, respectively, for detecting weak associations,^[3] detecting and correcting the bias of unmeasured factors,^[4] and testing treatment effects in randomized controlled trials.^[7] Along this line of inquiry, in this paper we propose a new avenue for health outcome prediction.

The C1 condition is the fundamental assumption of our method; it stipulates that the natural course of the health outcome of concern should produce perturbations as such. The C1 condition also implies that there is an intermediate state between the absence and the presence of the outcome: the outcome-developing state. For some health outcome that develops very quickly, the C1 condition may not apply. The C2 condition signifies that collecting numerous attributes does not always help; only the perturbed attributes count as the informative “signals” and the non-perturbed attributes are the uninformative “noises.” To meet this condition, as the number of attributes tends to infinity, the proportion of the perturbed attributes (signal prevalence) must be no smaller than a certain positive value. The C3 condition requires that the perturbations should be non-negligible if small; the perturbation magnitude (signal strength)—as measured by the deviation of the mean of those developing the outcome from that of those without the outcome, relative to the distance between the high and low limits set up for that attribute—must be no smaller than a certain positive value. Finally, the C4 condition stipulates that diverse types of attributes must be collected to minimize the correlations between them; as the number of attributes tends to infinity, the number of other attributes that an attribute correlates with must be bounded.

The current version of the PUMP only accepts binary attributes. Additional work is needed to expand the range of applicability to include categorical/continuous attributes, and those attributes that may change over time either because of the outcome-developing processes or by their own nature. The PUMP is also very naive; it does not care to differentiate signals

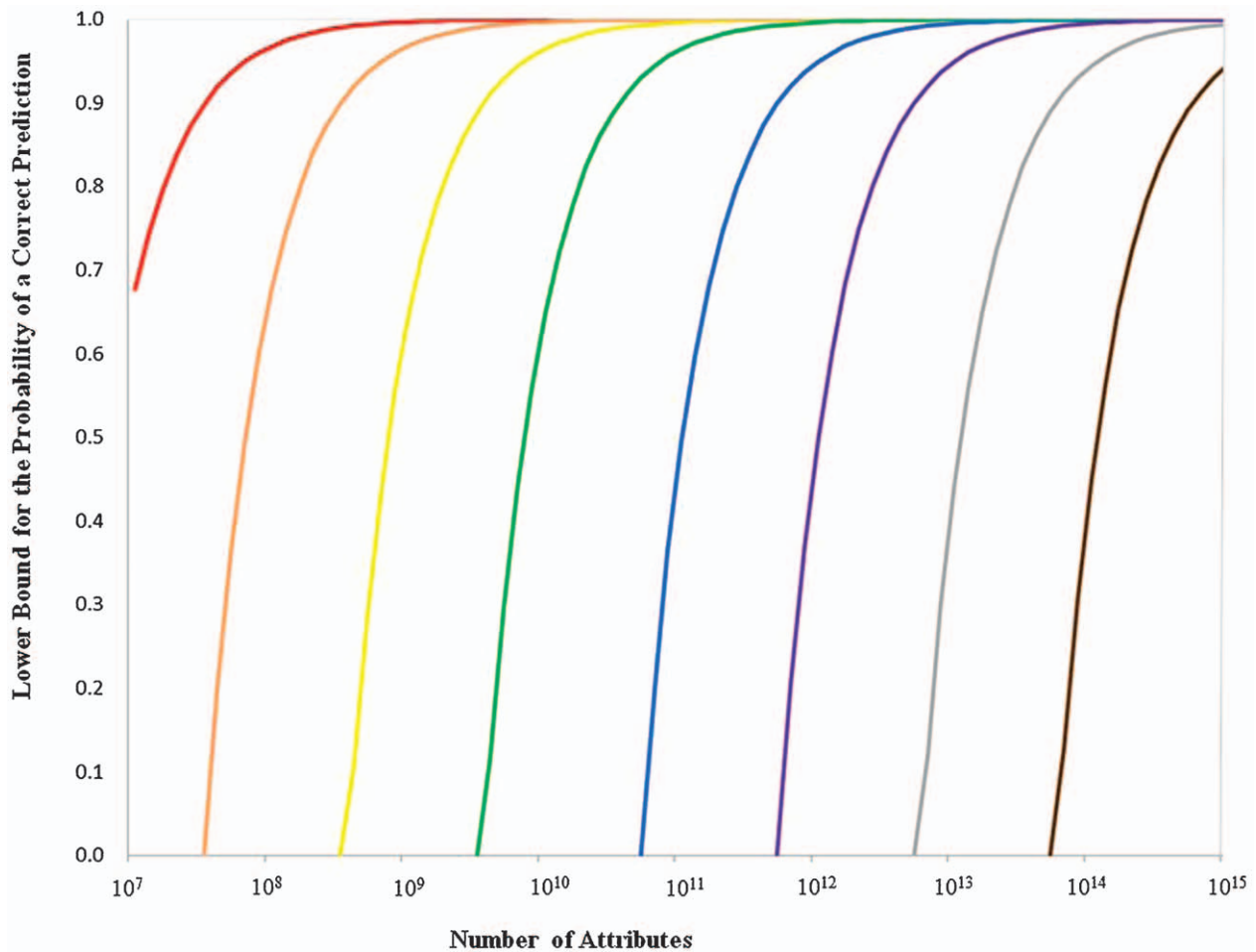


Figure 1. Lower bound for the probability of a correct prediction (red: corresponding to Scenario I in Table 1; yellow: Scenario II; blue: Scenario III; gray: Scenario IV; orange: Scenario V; green: Scenario VI; purple: Scenario VII; brown: Scenario VIII).

(the perturbed attributes) from noises (the non-perturbed attributes) before taking them all in. Methods to select attributes for the PUMP need further development. From an artificial intelligence perspective, the PUMP is the simplest “machine learner” which takes in one layer of attributes, performs a simple linear combination of them, and outputs an average perturbation score. Recent “deep learners” allow multiple processing layers

and complex nonlinear combinations between the input attributes and each and every node in the deep layers.^[8] Further studies along this line are also warranted. From Table 1, we see that the current version of the PUMP requires an extremely large number of attributes. Future updates of the PUMP envisioned above may reduce the number of attributes needed for a near perfect prediction to a practically feasible level.

Table 1

Numbers of attributes needed to control a false positive rate (the probability of a wrong prediction for a person without the outcome) and a false negative rate (the probability of a wrong prediction for a person developing the outcome) both no larger than 0.01 under various scenarios.

Scenario	Maximum number of other attributes an attribute correlates with	Proportion of perturbed attributes	Deviation of the mean of those developing the outcome from that of those without the outcome	Sample sizes of the training data			Numbers of attributes needed
				Case sample	1st control sample	2nd control sample	
I	100	0.05	0.05	1000	1000	1000	3.2×10^8
II	100	0.05	0.05	10	10	10	3.6×10^{10}
III	100	0.01	0.01	1000	1000	1000	5.1×10^{12}
IV	100	0.01	0.01	10	10	10	5.6×10^{14}
V	1000	0.05	0.05	1000	1000	1000	3.2×10^9
VI	1000	0.05	0.05	10	10	10	3.5×10^{11}
VII	1000	0.01	0.01	1000	1000	1000	5.0×10^{13}
VIII	1000	0.01	0.01	10	10	10	5.5×10^{15}

At present, it is still not possible to collect billions or trillions of attributes of a person. But moving into this big data era,^[9–11] we are gradually closing that gap. This paper shows that if such personal big data can be obtained and the C1–C4 conditions are met, theoretically we can use a PUMP to predict the health outcome of a person to near certainty.

Author contributions

This is a single-authorship paper by WCL.

Conceptualization: Wen-Chung Lee.

Data curation: Wen-Chung Lee.

Formal analysis: Wen-Chung Lee.

Funding acquisition: Wen-Chung Lee.

Investigation: Wen-Chung Lee.

Methodology: Wen-Chung Lee.

Project administration: Wen-Chung Lee.

Resources: Wen-Chung Lee.

Software: Wen-Chung Lee.

Supervision: Wen-Chung Lee.

Validation: Wen-Chung Lee.

Visualization: Wen-Chung Lee.

Writing – original draft: Wen-Chung Lee.

Writing – review & editing: Wen-Chung Lee.

References

- [1] Wilson PWF, D’Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97:1837–47.
- [2] Benedet JL, Bender H, Jones H, et al. FIGO staging classifications and clinical practice guidelines in the management of gynecologic cancers. *Int J Gynaecol Obstet* 2000;70:209–62.
- [3] Lo MT, Lee WC. Detecting a weak association by testing its multiple perturbations: a data mining approach. *Sci Rep* 2014;4:5081.
- [4] Lee WC. Detecting and correcting the bias of unmeasured factors using perturbation analysis: a data-mining approach. *BMC Med Res Methodol* 2014;14:18.
- [5] Hall P, Marron JS, Neeman A. Geometric representation of high dimension, low sample size data. *J Royal Stat Soc (series B)* 2005;67:427–44.
- [6] Ahn J, Marron JS, Muller KM, et al. The high-dimension, low-sample-size geometric representation hold under mild conditions. *Biometrika* 2007;94:760–6.
- [7] Lee WC, Lin JH. A test for treatment effects in randomized controlled trials, harnessing the power of ultrahigh dimensional big data. *Medicine* 2019;98:43(e17630).
- [8] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
- [9] Murdoch TB, Detsky AS. The inevitable application of big data to health care. *J Am Med Assoc* 2013;309:1351–2.
- [10] Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff (Millwood)* 2014;33:1163–70.
- [11] Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics* 2015;8:33.