Received: 15 January 2019 | Revised: 11 September 2019 | Accepted: 18 September 2019

DOI: 10.1002/bimj.201900046

Biometrical Journal

RESEARCH PAPER

# Berkson's paradox and weighted distributions: An application to Alzheimer's disease

**Polychronis Economou**[1] | **Apostolos Batsidis**[2] | **George Tzavelas**[3] | **Panagiotis Alexopoulos**[4,5] | **Alzheimer's Disease Neuroimaging Initiative**

[1]Department of Civil Engineering, University of Patras, Rion-Patras, Greece

[2]Department of Mathematics, University of Ioannina, Ioannina, Greece

[3]Department of Statistics and Insurance Science, University of Piraeus, Piraeus, Greece

[4]Department of Psychiatry, Faculty of Medicine, University of Patras, University Hospital of Rion, Rion Patras, Greece

[5]Department of Psychiatry and Psychotherapy, Faculty of Medicine, Technical University of Munich, Klinikum rechts der Isar, Munich, Germany

**Correspondence**
Polychronis Economou, Department of Civil Engineering, University of Patras, 265 00, Rion-Patras, Greece.
Email: peconom@upatras.gr

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

**Abstract**

One reason for observing in practice a false positive or negative correlation between two random variables, which are either not correlated or correlated with a different direction, is the overrepresentation in the sample of individuals satisfying specific properties. In 1946, Berkson first illustrated the presence of a false correlation due to this last reason, which is known as Berkson's paradox and is one of the most famous paradox in probability and statistics. In this paper, the concept of weighted distributions is utilized to describe Berskon's paradox. Moreover, a proper procedure is suggested to make inference for the population given a biased sample which possesses all the characteristics of Berkson's paradox. A real data application for patients with dementia due to Alzheimer's disease demonstrates that the proposed method reveals characteristics of the population that are masked by the sampling procedure.

**KEYWORDS**

ABC rejection algorithm, Alzheimer's disease, Berkson's fallacy, biased sampling, likelihood-free inference

## 1 | INTRODUCTION

Very often in practical applications a false (positive or negative) correlation is observed between two random variables (r.v.), say $X$ and $Y$, that are either not correlated or correlated with a different direction. The observed false correlation may occur due to a number of reasons. For instance, there might be a third, lurking variable, say $Z$, which is not included in the study or is difficult to be identified, that makes the relationship appear stronger or weaker than it actually is. Another reason which can lead to observe a false correlation among two characteristics is the overrepresentation in the sample of individuals satisfying specific properties. Berkson (1946) first illustrated the presence of a false correlation due to this last reason, with a case-control

  

study linking diabetes with cholecystitis amongst inpatients who seek care. The two diseases were found to be positive correlated even if they are independent in the population. Berkson himself explained this spuriously finding by recognizing that a patient with more than one disease was more likely to be hospitalized than a patient with only a single disease. Since then, such false observed correlation due to a bias sample, is known as Berkson's paradox or Berkson's bias or fallacy. It can arise in prospective or retrospective studies, and in randomized or observational settings.

Berkson's paradox is widely recognized in many fields including medicine and epidemiology (Feinstein, Walter, & Horwitz, 1986; Peritz, 1984; Westreich, 2012) and social sciences (Morgan, 2013). Some authors encourage physicians to understand Berkson's paradox in order to avoid misinterpreting data whenever counter-intuitive findings are observed and some others have considered aspects of adjusting these findings. For instance, Geneletti, Best, Toledano, Elliot, and Richardson (2013) introduced the so-called Bias Breaking Model to provide a statistical solution to selection bias. The basic assumption behind this model is that there is a variable, termed the bias breaking variable, which is associated with both the selection and the exposure, and in some way separates them. The conditions for a variable to be bias breaking are formulated in terms of conditional independencies and are represented by directed acyclic graphs (DAGs). However, in many cases, it is difficult to identify the bias breaking variable. The structure of bias due to the selection procedure was first described in the DAG literature by Pearl (1995) and Spirtes, Glymour, and Scheines (1993). For more details on causal diagram theory and selection bias, see for instance Greenland (2003); Hernan, Hernandez-Diaz, and Robins (2004); VanderWeel, Herman, and Robins (2008). Some other methods used to adjust selection bias are the poststratification (Samuelsen et al., 2007) and the inverse probability weighting (IPW) (Rotnitzky & Robins, 2005). Both methods are aimed at adjusting for potential biases. Poststratification is used to achieve better accuracy by making the sample more representative of the target population, while IPW uses external data to assign to each subject a weight which is the inverse of its probability of selection.

Since Berkson's paradox occurs actually by the overrepresentation in the sample of individuals satisfying specific properties, the Berkson's paradox can be interpreted as a selection bias problem, resulting to a sample which is not representative of the population intended to analyzed. Extending the basic ideas of Fisher (1934), Rao (1965) introduced the concept of a weighted distribution as a method of adjustment applicable to many situations in which the recorded observations cannot be considered as a random sample from the original distribution due for instance to the nonobservability of some events. The purpose of the present paper is to propose a method, using the concept of weighted distributions, to make inference for the population given a biased sample which possesses all the characteristics of Berkson's paradox. Although, the proposed method shares some ideas with the aforementioned IPW method, it has to be pointed out that the IPW applies, to the best of our knowledge, to $2 \times 2$ tables and demands external data. On the other hand, the proposed method, based on the weighted distributions, applies to data from continuous r.v.'s under the assumption that the parametric form of the distribution of the r.v.'s is known. Finally, comparing with methods for adjusting selection bias like the Bias Breaking Model there is no need for identifying a third variable which is associated with the selection.

The rest of the paper is organized as follows. In Section 2, a brief introduction to the concept of weighted distributions is given and weight functions are proposed which can formulate four different scenarios, in which members of the population are overrepresented in the sample and this overrepresentation causes falsely observed correlations. Because in most cases the likelihood of the proposed model is complex, the approximate Bayesian computation rejection algorithm, a likelihood-free method, is used for statistical inference in Section 3. Some necessary details for implementing this algorithm in practice are also given in Section 3. In Section 4, a real data application for patients with dementia due to Alzheimer's disease is presented. Finally, some concluding remarks are given in the last section.

## 2 | WEIGHT FUNCTIONS FOR MODELING BERKSON'S PARADOX

As already mentioned, Berkson's paradox is actually a selection bias problem, caused from the fact that the sample being observed is not a random sample of the population. Rao (1965) introduced the concept of the univariate weighted distributions to describe the biasness in a sample. Biased samples do not only emerge by applying unintentionally a biased sampling scheme but in many cases arise naturally by the nature of the problem (see, e.g., Afonso & Corte Real, 2016; Gupta & Kirmani, 1990; Patil & Rao, 1978). Moving one step forward, we will adopt the following definition of bivariate weighted distributions (see for instance Mahfoud & Patil, 1982; Sarabia & Gomez-Deniz, 2008) for the study of Berkson's paradox in order to adjust the bias in the sample.

**Definition 2.1.** Let $(X, Y)$ be a two-dimensional random vector with joint probability density function (p.d.f.) $f(x, y; \theta)$, where $\theta$ is an unknown $s$-dimensional parameter, which belongs on a parameter space $\Theta$, where $\Theta \subseteq R^s$ with $s \geq 1$. The two-dimensional

random vector $(X_w, Y_w)$ with joint p.d.f.

$$f_w(x, y; \theta) = \frac{w(x, y; \theta)}{E\{w(X, Y; \theta)\}} f(x, y; \theta), \tag{1}$$

is called the vector of the weighted random variables corresponding to $(X, Y)$, associated with $w(x, y; \theta)$, a non-negative function, that is, $w : R^2 \to R^+$, such that $E\{w(X, Y; \theta)\} < \infty$. The p.d.f. $f_w(x, y; \theta)$ is called the bivariate weighted pdf corresponding to $f(x, y; \theta)$, associated with $w$, while any random vector with density $f_w$ is denoted by $(X_w, Y_w)$.

Some properties of the bivariate and multivariate weighted distributions can be found, among others, in Arnold and Nagaraja (1991), Jain and Nanda (1995), Nanda and Jain (1999) and Navarro, Ruiz, and Aguila (2006). Arnold and Nagaraja (1991) focused on the dependence in bivariate weighted distributions and gave the following result (for a proof, see Alavi & Chinipardaz, 2007).

**Proposition 2.2** (Arnold & Nagaraja, 1991)**.** *For an arbitrary weight function $w(x, y)$ any two of the following statements together imply the third:*

*i) $X$ and $Y$ are independent,*

*ii) $X_w$ and $Y_w$ are independent,*

*iii) $w(x, y)$ is of the form $w_1(x)w_2(y)$ for $(x, y) \in S_1 \times S_2$, the Cartesian product of the supports of $X$ and $Y$, respectively.*

Based on Proposition 2.2, if $X$ and $Y$ are independent r.v.s, then, in order for $X_w$ and $Y_w$ to be dependent, the weight function should not be of the form $w_1(x)w_2(y)$. One possible family of weight functions, among many others, that induce dependence among the weight r.v.s may be given by

$$w(x, y; \theta, \gamma_1, \gamma_2) = 1 - h(x; \theta, \gamma_1)g(y; \theta, \gamma_2) \tag{2}$$

for suitable choices $h(x; \theta, \gamma_1)$ and $g(y; \theta, \gamma_2)$, where $\theta$ is the s-dimensional vector of parameters of the bivariate distribution of $(X, Y)$ and $\gamma = (\gamma_1, \gamma_2)$ is a vector of extra parameters with $\gamma_1, \gamma_2 > 0$. Note that $h(x; \theta, \gamma_1)$ and $g(y; \theta, \gamma_2)$ should guarantee that the weight function $w(x, y; \theta, \gamma_1, \gamma_2)$ is non-negative, with $E\{w(X, Y; \theta, \gamma_1, \gamma_2)\} < \infty$. At the same time both functions should be such that the $w(x, y; \theta, \gamma_1, \gamma_2)$ to reflect the probability of a pair $(x, y)$ to be selected in the sample. This means for instance that $h(x; \theta, \gamma_1)$ and $g(y; \theta, \gamma_2)$ should take small (large) values in areas in which a pair $(x, y)$ has high (small) probability to be selected in the sample.

In the sequel, four general cases of a biased sample which may cause Berkson's paradox will be studied. Let us consider first the case (Case A hereafter) where units with large values on $X$ and/or large values on $Y$ are more likely to be selected. This means that the nonrandom sampling mechanism gives to pairs $(x, y)$ with large values on $X$ and/or on $Y$ high probability to be observed. On the other hand, it gives small probability to pairs $(x, y)$ with small values on both $X$ and $Y$. One reasonable and mathematically convenient choice, among an infinite number of functions with the above behavior, for the weight function $w(x, y; \theta, \gamma_1, \gamma_2)$ may be the following:

$$w_1(x, y; \theta, \gamma_1, \gamma_2) = 1 - \{1 - F_X(x; \theta_x)\}^{1/\gamma_1}\{1 - F_Y(y; \theta_y)\}^{1/\gamma_2}, \text{ for } \gamma_1, \gamma_2 \geq 0, \tag{3}$$

where $\theta_x$ and $\theta_y$ are functions of $\theta$ and $F_X(x; \theta_x)$ and $F_Y(y; \theta_y)$ are the c.d.f. of $X$ and $Y$, respectively.

Case A is illustrated in the first plot in Figure 1. In this plot, a simulated sample of size $n = 50$ (denoted by black points) was generated from the bivariate random vector $(X_{w_1}, Y_{w_1})$ corresponding to the random vector $(X, Y)$, with $X$ and $Y$ independent normal distributed r.v.s. with mean zero and standard deviation 2 and with values of $\gamma_1, \gamma_2$ similar to those obtained by the data set used in Section 4. For comparison purposes the contour plot of the p.d.f. of $(X, Y)$ is also given. From the plot it is obvious that there is indeed a tendency to observe pairs with large values on $X$ and/or $Y$ and that the left bottom part of the population is underrepresented in the sample.

Except from Case A, three other types of sampling mechanisms which may cause false correlation, due to the biasness in the sample, will be discussed. Case B corresponds to the nonrandom sampling where units with small values on $X$ and/or small values on $Y$ are more likely to be observed. In this case, with similar arguments, one reasonable and mathematically convenient choice for the weight function $w(x, y; \theta, \gamma_1, \gamma_2)$ is the following:

$$w_2(x, y; \theta, \gamma_1, \gamma_2) = 1 - \{F_X(x; \theta_x)\}^{1/\gamma_1}\{F_Y(y; \theta_y)\}^{1/\gamma_2}. \tag{4}$$

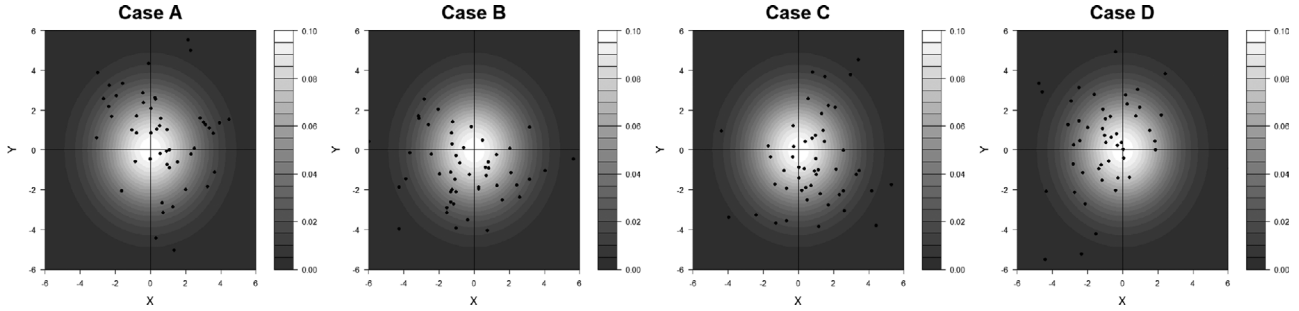


**FIGURE 1** The contour plot of independent normal distributed random variables $(X, Y)$ with mean zero and standard deviation 2 and the simulated samples of size $n = 50$ (black points) for the four cases of Berkson's paradox. Cases A (first plot) and B (second plot) illustrate a false negative correlation while cases C (third plot) and D (fourth plot) illustrate a false positive correlation

In Case C (Case D) pairs $(x, y)$ with large (small) values on $X$ and/or small (large) values on $Y$ are more likely to be included in the sample. In this frame, the following weight functions:

$$w_3(x, y; \theta, \gamma_1, \gamma_2) = 1 - \{1 - F_X(x; \theta_x)\}^{1/\gamma_1} \{F_Y(y; \theta_y)\}^{1/\gamma_2}, \tag{5}$$

and

$$w_4(x, y; \theta, \gamma_1, \gamma_2) = 1 - \{F_X(x; \theta_x)\}^{1/\gamma_1} \{1 - F_Y(y; \theta_y)\}^{1/\gamma_2}, \tag{6}$$

are reasonable choices for Case C and D, respectively.

Cases B, C and D are illustrated in the second, third and fourth plot in Figure 1. The observed samples of size $n = 50$ were generated by assuming a bivariate normal distribution with the same characteristics as previously; however, the weight functions given in relations (4), (5), and (6) were used.

*Remark* 2.3. As already mentioned, the weight functions $w_i(x, y; \theta, \gamma_1, \gamma_2)$, $i = 1, \ldots, 4$, are just reasonable functions to be adopted under the concept of weighted distributions in order to model Cases A–D. Of course, there are infinity other possible forms of weight functions that may also be used to model such situations, although, we emphasize to the convenient monotonicity property with respect to $x$ and $y$ that these weight function posses and to the use of the marginal c.d.f. of $X$ and $Y$ which make these functions a more instinctive choice than others.

In the sequel, the effect of the weight functions $w_i(x, y; \theta, \gamma_1, \gamma_2)$, $i = 1, \ldots, 4$, on $cov(X_w, Y_w)$, that is, on the dependence structure of $X_w$ and $Y_w$ will be studied. The properties obtained will be later on used in order to help us choose one of the four weight functions for real applicative purposes. Before we proceed, some useful definitions are presented.

**Definition 2.4.** We say that a non-negative function $\phi(x, y)$ is a reverse regular of order 2 (RR2) if $\phi(x_1, y_1)\phi(x_2, y_2) \leq \phi(x_1, y_2)\phi(x_2, y_1)$ whenever $x_1 < x_2$ and $y_1 < y_2$, while we say that is totally positive of order 2 (TP2) if the inequality is reversed.

**Proposition 2.5.** *Suppose that $X$ and $Y$ are independent r.v.s. Then $cov(X_{w_i}, Y_{w_i}) \leq 0$, for $i = 1, 2$, while $cov(X_{w_i}, Y_{w_i}) \geq 0$, for $i = 3, 4$.*

*Proof.* It is easily proved that the weight function $w_i(x, y; \theta, \gamma_1, \gamma_2)$ is RR2 for $i = 1, 2$, while $w_i(x, y; \theta, \gamma_1, \gamma_2)$ is TP2 for $i = 3, 4$. Hence from Theorems 4.1. and 4.2 given in Nanda and Jain (1999), it is concluded that $cov(X_{w_i}, Y_{w_i}) \leq 0$, for $i = 1, 2$, while $cov(X_{w_i}, Y_{w_i}) \geq 0$, for $i = 3, 4$. □

In the previous proposition, the effect of the weight functions $w_i(x, y; \theta, \gamma_1, \gamma_2)$, $i = 1, \ldots, 4$, on the dependence structure of $X_{w_i}$ and $Y_{w_i}$, was presented under the assumption of the independence of $X$ and $Y$. The next proposition states the preservation property of some dependence concepts under the weighting used in this paper. In the sequel, PLRD stands for positive likelihood ratio dependence, while NLRD stands for negative likelihood ratio dependence. Note that when $X$ has a density $f$ then PLRD (NLRD) is equivalent to the condition that $f$ is TP2 (RR2). For more definitions and details of the dependence concepts used in the sequel, we refer the reader to Nelsen (2006).

**Proposition 2.6.**

*(i) Let* $(X,Y)$ *be PLRD; then* $(X_{w_3},Y_{w_3})$ *and* $(X_{w_4},Y_{w_4})$ *are PLRD, which implies that* $cov(X_{w_3},Y_{w_3})\geq 0$ *and* $cov(X_{w_4},Y_{w_4})\geq 0$.

*(ii) Let* $(X,Y)$ *be NLRD; then* $(X_{w_1},Y_{w_1})$ *and* $(X_{w_2},Y_{w_2})$ *are NLRD, which implies that* $cov(X_{w_1},Y_{w_1})\leq 0$ *and* $cov(X_{w_2},Y_{w_2})\leq 0$.

*Proof.*

(i) After some algebra and based on Izadkhah, Amini, and Borzadaran Mohtashami (2016, Theorem 1), when $(X,Y)$ is PLRD, then $(X_{w_3},Y_{w_3})$ and $(X_{w_4},Y_{w_4})$ are PLRD, which implies $cov(X_{w_3},Y_{w_3})\geq 0$ and $cov(X_{w_4},Y_{w_4})\geq 0$.

(ii) It is easily proved that the weight functions $w_1(x,y;\theta,\gamma_1,\gamma_2)$ and $w_2(x,y;\theta,\gamma_1,\gamma_2)$ are reverse regular of order 2 (RR2) in $(x,y)$. Then based on Izadkhah, Amini, and Borzadaran Mohtashami (2016, Theorem 1), when $(X,Y)$ is NLRD then $(X_{w_1},Y_{w_1})$ and $(X_{w_2},Y_{w_2})$ are NLRD, which implies $cov(X_{w_1},Y_{w_1})\leq 0$ and $cov(X_{w_2},Y_{w_2})\leq 0$. □

Proposition 2.6 states that the covariance of $(X,Y)$ and the covariance of its weighted version $(X_{w_i},Y_{w_i})$, $i=3,4$ $((X_{w_i},Y_{w_i})$, $i=1,2)$ will always have the same sign or will be zero whenever $(X,Y)$ is PLRD (NLRD). For instance, if $(X,Y)\sim N_2(\mu,\Sigma)$ with correlation coefficient $\rho\geq 0$ then $(X,Y)$ is PLRD and the weight functions $w_3$ and $w_4$ ensure that $cov(X_{w_i},Y_{w_i})\geq 0$, for $i=3,4$. On the other hand, if $(X,Y)\sim N_2(\mu,\Sigma)$ with $\rho\leq 0$ then $(X,Y)$ is NLRD and the weight functions $w_1$ and $w_2$ ensure that $cov(X_{w_i},Y_{w_i})\leq 0$, for $i=1,2$. However, it is an open problem to make a general comment about the sign of $cov(X_{w_i},Y_{w_i})$, $i=1,\ldots,4$ as well as a comment about the sign of the difference between $cov(X_{w_i},Y_{w_i})$, $i=1,\ldots,4$ and $cov(X,Y)$, without the previous assumptions.

Before closing this section, taking into account the interpretation of the four weight functions and the results in Propositions 2.5 and 2.6, practical rules in order to choose between one of the four weight functions for real applicative purposes are given in the next remark, while an example of how this choice can be made in a real life problem is presented in details in Section 4.

*Remark* 2.7. The choice of the weight function is mainly based on the nature of the problem and the characteristics inherited to the observed sample by each weight function. If the nonrandom sampling is known, then $w_1$ ($w_2$) is adopted when units with large (small) values on $X$ and/or large (small) values on $Y$ are more likely to be selected, while $w_3$ ($w_4$) is adopted when units with large (small) values on $X$ and/or small (large) values on $Y$ are more likely to be selected. Consider now the case that no prior information about the nonrandom sampling is available; however, it is beforehand known that $X$ and $Y$ are independent. Then, based on Proposition 2.5, in case of false negative correlation one has to choose between the weight functions $w_i(x,y;\theta,\gamma_1,\gamma_2)$, $i=1,2$, since $cov(X_{w_i},Y_{w_i})<0$, $i=1,2$. On the other hand, if a false positive correlation is observed, then the two candidate weight functions are $w_i(x,y;\theta,\gamma_1,\gamma_2)$, $i=3,4$, since $cov(X_{w_i},Y_{w_i})>0$, $i=3,4$. To choose, in both cases, between the two candidate weight functions, one should again examine the nature of the sampling mechanism and recognize which pairs are more likely to be observed and which not, as previously explained.

Finally, if $X$ and $Y$ are not independent, but they are PLRD (NLRD) then, based on Proposition 2.6, the weight functions $w_i$, $i=3,4$ ($w_i$, $i=1,2$) ensure that the covariance of the weighted variables will have the same sign or will be zero. Therefore, if $X$ and $Y$ are PLRD (NLRD) then $w_i$, $i=3,4$ ($w_i$, $i=1,2$) cannot be used for modeling Berkson's paradox, that is, for modeling false negative (positive) correlation. Next, again the selection between the other two candidate weight functions relies on the correct recognition of the sampling mechanism nature as before and with the rules given previously.

## 3 | INFERENCE UNDER THE PROPOSED MODEL

The likelihood function, given a biased sample of $n$ observations $D_i=(X_i,Y_i)$, $i=1,\ldots,n$ from a parent population with known, based on previous studies, parametrical form, but with unknown parameters, is in most cases complex. As a result, evaluating the likelihood can be computationally very expensive and in some cases even impossible to write it analytically. These characteristics make the adoption of a likelihood-free method (Diggle & Gratton, 1984; Rubin, 1984) necessary to perform inference. Such a method is the Approximate Bayesian computation (ABC) method which constitutes a class of computational methods routed in Bayesian Statistics. The ABC rejection algorithm is described in the following steps:

1. a) Adopt a prior distribution for each parameter involved in the expression of the weighted distribution, that is, for each component of the vector of parameters $\theta$ of the bivariate random variable and $\gamma_1, \gamma_2$.
   b) Simulate $\theta^*$, $\gamma_1^*$, and $\gamma_2^*$ from the previous prior distributions. Denote $\zeta^* = (\theta^*, \gamma_1^*, \gamma_2^*)$.
2. Plug $\zeta^*$ to the proposed weighted distribution and simulate a sample $D_i^* = (X_i^*, Y_i^*)$, $i = 1, \ldots, n$ from the model with p.d.f. $f_w(x, y; \theta^*, \gamma_1^*, \gamma_2^*)$.
3. Compute the discrepancy between the simulated and the observed data set, that is, compute $\Delta_{\zeta^*} = d(D, D^*)$, where $d$ is a discrepancy measure.
4. Retain $\zeta^*$ if $\Delta_{\zeta^*} < \epsilon$, for some $\epsilon > 0$. Go to step 1 b) and repeat $M$ times.

The above algorithm describes the procedure to obtain a sample from a distribution close to the posterior distribution of $\zeta^*$ by comparing simulated samples with the original sample. In the sequel some necessary details for implementing the previous algorithm in practice are discussed.

1. A prior distribution for each component of the parameter $\theta$ as well as for $\gamma_1$ and $\gamma_2$ should be selected. Any previous knowledge has to be implemented in these distributions. Fortunately, such knowledge for the components of $\theta$ is expected to exist, since we are aware of possible occurrence of Berkson's paradox. On the other hand, no prior information is expected to be available for the $\gamma_1, \gamma_2$ parameters. For that reason, initially prior distributions with large standard deviations are recommended. After a number of nonrejections (see Step 4), the prior distributions of all the parameters can be updated using the information gained from these samples to increase the rate of acceptance.

   To keep the proposed method as simple as possible we recommend the use of truncated at zero normal distribution as prior distribution for any positive parameter, the normal distribution for any real parameter and the uniform distribution for any parameter bounded on a finite interval [a,b]. These prior distributions are the maximum entropy priors given only the continuity and support of the parameters (Schroeder, 2010). Such priors may not be the optimal, since any other prior distribution that incorporates in a better, more informative way, any prior expert knowledge from the problem domain is of course preferable. For instance, in the application presented in Section 4, a linear transformation of a beta distribution is adopted as a prior for the correlation coefficient. If nonoptimal priors are used, the method can still yield reasonable parameter estimates, although adopting such priors may result in an increased rejection rate and, therefore, to a larger number of simulations (see Step 4).
2. One should acknowledge that simulating a random sample directly from $f_w(x, y; \zeta^*)$ is not always an easy task. Fortunately, in most of the cases, a large number, $N$, of observations can be generated from $f(x, y; \theta^*)$. This sample of $N$ observations can serve as the "population." To obtain a sample from $f_w(x, y; \zeta^*)$, one can then apply a weighted sampling without replacement procedure to select $n$ observations from the "population" with weights proportional to $w(x, y; \zeta^*)$. Obviously, this procedure adds another layer of approximation to the posterior distribution of the parameters but this should not be that significant if the "population" size $N$ is relatively large compared to the sample size $n$.
3. A discrepancy measure can be defined using the integrated squared error given by

$$T = \int_x \int_y \left( f_D(x, y) - f_{D^*}(x, y) \right)^2 dy dx,$$

   where $f_D(x, y)$ and $f_{D^*}(x, y)$ are the kernel density estimates based on the observed $D$ and the simulated $D^*$ sample, respectively. Under the (null) hypothesis that both samples share the same density, $T$ is asymptotically normally distributed with known mean and standard deviation. As a consequence, the absolute value of the standardized version of $T$ can serve as discrepancy measure $d(D, D^*)$ between the two samples (for more details, see Duong, Goud, & Schauer, 2012).
4. Values of $\epsilon$ close to 0 guarantee that the retained values of $\zeta^*$ consist of a sample from a distribution close to the posterior distribution of $\zeta$. However, very small values lead to high rejection rates and therefore to a large number of simulations to obtain a relative large number of observations from the posterior distribution. We recommend the use of $\epsilon = 1.96$ that corresponds to a significant level 0.05 for testing the null hypothesis that both samples share the same density.

*Remark* 3.1. ABC rejection algorithm allows the handling of complex models with a large number of parameters, although as the number of parameters increases, the acceptance rate of the simulated samples decreases exponentially due to the global acceptance criterion (Csilléry, Blum, Gaggiotti, & François, 2010). As a consequence, complex models require large number of simulations that may be time consuming or can require significant computing power.

# 4 | APPLICATION

## 4.1 | Alzheimer's disease

Brain hypometabolism is related to both Alzheimer's disease (AD), the most common cause of late-onset dementia (Alzheimer's Association, 2016; Rasmussen, Tybjæaerg-Hansen, Nordestgaard, & Frikke-Schmidt, 2018) and normal aging. AD is characterized by a reduction of metabolism in bilateral angular gyrus, posterior cingulate/precuneus and inferior temporal cortex (Rice & Bisdas, 2017). Fluorine-18-labeled fluorodeoxyglucose ($^{18}$F-FDG), a radioligand used in positron emission tomography (PET), enables the detection of the typical AD hypometabolic pattern not only in AD dementia, but also in mild cognitive impairment (MCI) due to AD, a predementia stage of AD (Alexopoulos, Grimmer, Perneczky, Domes, & Kurz, 2006). Moreover, $^{18}$F-FDG uptake decreases in large regions of the brain with advancing age in cognitively normal elderly individuals with or without AD neuropathological changes (Ishibashi et al., 2018; Jiang et al., 2018; Jack et al., 2018).

The application is based on data from the AD Neuroimaging Initiative (ADNI) databank. ADNI is a collaboration between approximately 50 academic institutions and private corporations in the USA and Canada and is supported by the National Institute on Aging (NIA), nonprofit organizations and private pharmaceutical companies. General eligibility criteria are described at www.adni-info.org/Scientists/ADNIGrant/ProtocolSummary.aspx. The study procedures were approved by the institutional review boards of all participating centers and written informed consent was obtained from all participants or authorized representatives. All procedures performed in the study were in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The data for this study were obtained from www.adni-info.org in 2014 (Robb et al., 2017) and consists of $n = 76$ patients with AD dementia.

The purpose of this application is to examine if the proposed method can reveal characteristics of the population of interest based on a biased sample. More specifically, we will restrict our data to the data set on which normally a clinic-based study for patients with AD will be mainly focused, that is, patients with AD dementia. The population of interest encompasses individuals with memory concerns, that is, patients with AD dementia or with MCI and individuals with subjective memory complaints, who seek advice and care at memory clinics. Obviously, this population does not include all the elderly individuals but reflects the whole continuum of cognitive performance in aging, ranging from very mild cognitive deterioration due to healthy cognitive aging to dementia.

*Remark* 4.1. Before proceeding in the analysis based on the weight functions proposed in Section 2, we have to note once again that the weight functions used is only some possible and logical choices from a collection of an arbitrary number of weight functions, which fulfilled the same properties. Moreover, as a referee pointed out, the proposed model is just one among other possible techniques (see also a remark in the next section) that can be used to deal with Berkson's paradox.

## 4.2 | Sampling procedure and the proposed method

For the patients with AD dementia included in the present application, we have recorded their age $X$ (as an important risk factor for the development of AD) and their FDG PET scores $Y$. The distribution of both FDG PET scores and age in the population of interest can be approximated by a normal distribution. As a result, we assume that the population distribution of the bivariate random vector $(X, Y)$ is the bivariate normal distribution with location parameter $(\mu_x, \mu_y)'$, scale parameters $\sigma_x > 0$ and $\sigma_y > 0$ and $\rho$ the correlation between $X$ and $Y$, which will actually be the main interest of the present analysis. Based on prior knowledge $\rho$ is expected to be nonpositive ($\rho \leq 0$) due to the negative influence of advancing age on brain metabolism. Regarding the age, it is of note that the total number of people with dementia, MCI and subjective memory complaints, follows a unimodal distribution which rises up to the age of 80–84 years and afterward declines (Drew, 2018; Fritsch, McClendon, Wallendal, Hyde, & Larsen, 2014; Prince et al., 2014; Warda, Arrighib, Michelsa, & Cedarbaum, 2011). This distribution, which approximates to normal distribution, is attributed to the changing balance between the increasing age-specific prevalence of cognitive deterioration and the increasing mortality as age advances (Drew, 2018; Prince et al., 2014).

For the 76 AD dementia patients in the study, the mean age was $\bar{x} = 75.253$ years (sd = 8.588) and the mean FDG PET score was $\bar{y} = 1.0761$ (sd = 0.1211). The scatterplot of the observed data is presented in Figure 2. To avoid unnecessary repetition, we defer the discussion of the details of this figure to Section 4.3. At this point, we only emphasize that there is a clear positive correlation between age and FDG PET scores. In support of the latter observation, the Spearman's correlation coefficient $r$ is equal 0.392 ($p$-value = 0.0005) indicating indeed a clear positive correlation. This finding contradicts the well-known, established negative influence of advancing age on brain metabolism. This finding can only be explained in the context of Berkson's paradox and the biased sample used.

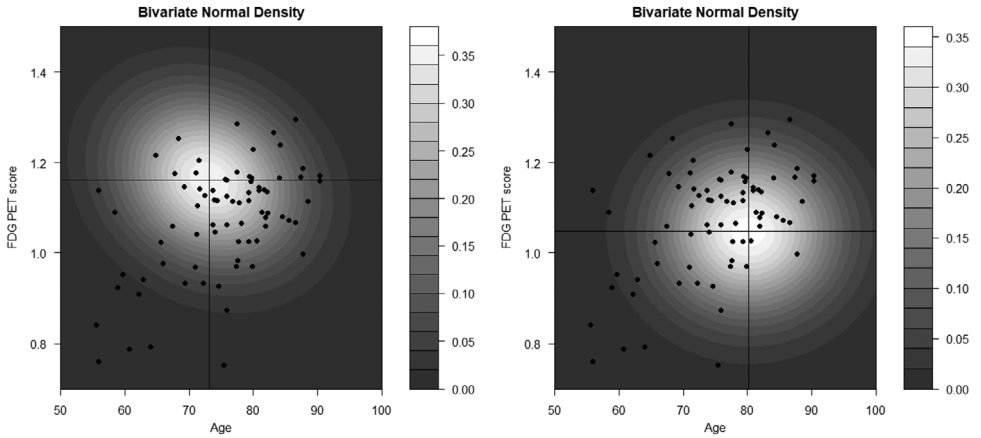


**FIGURE 2** The contour plot of the estimated bivariate normal (using the mean of the posterior distributions) under the incorrect $w_3$ (left plot) and the correct $w_4$ (right plot) weight function for the AD data

In order to incorporate in the analysis the bias caused by the sampling procedure, a weighted distribution of the form (1) can be fitted with $f_{XY}(x, y; \theta)$, the p.d.f. of the bivariate normal distribution. In the previous sections, four different weight functions were defined for dealing with four different cases of Berkson's paradox (Cases A–D). The next reasonable question is how to choose one of these four weight functions. Based on prior knowledge, it is established that $\rho \leq 0$, and hence, since $(X, Y)$ is assumed to be bivariate normal distributed, we have that $(X, Y)$ is NLRD. Taking into account Proposition 2.6 (ii), $(X_{w_1}, Y_{w_1})$ and $(X_{w_2}, Y_{w_2})$ is also NLRD, which implies that $cov(X_{w_1}, Y_{w_1}) \leq 0$ and $cov(X_{w_2}, Y_{w_2}) \leq 0$. Since in the biased sample, the Spearman's correlation coefficient $r$ is equal to 0.392 and indicates a clear positive correlation, the weight functions $w_1$ and $w_2$ are excluded as possible weight functions. Therefore, there are only two candidate weight functions, that is, the weight functions $w_3$ and $w_4$.

The next question is if there is a way to choose one of them based on the nature of the problem and the properties satisfied by each weight function. As it has already been underscored, advancing age pertains to both decrease in brain metabolism and increase in incidence of AD, which is characterized by brain hypometabolism. Nonetheless, participants of research initiatives, such as ADNI, who are highly sensible with regard to cognitive impairment, are not very old, so that they can serve as volunteers, and seek advice and help early, when they face only mild memory deficits and their cerebral hypometabolism is very mild too. As a result, a sample of a research initiative such as ADNI is more likely to include elderly individuals with smaller age and/or individuals with not so severe reduction in cerebral metabolism, that is, with relative high FDG PET scores. The previous properties are satisfied by the weight function $w_4(x, y;, \gamma_1, \gamma_2)$ (Case D, Figure 1). For comparison reasons although both $w_3(x, y;, \gamma_1, \gamma_2)$ (incorrect choice) and $w_4(x, y;, \gamma_1, \gamma_2)$ (correct choice) will be used to analyze the observed data.

*Remark* 4.2. As mentioned in the introduction, the false (positive in this application) correlation may occur due to a number of reasons. One such reason could be the presence of a third, lurking variable, say $Z$. For instance, a referee pointed out that the apparent positive correlation between $X$ and $Y$ in the observed sample could possibly be explained or even turned around with some third variable such as overall health status. Even if such an r.v. could be identified and included in the analysis, it may not offered any clear explanation on the overrepresentation in the sample of individual satisfying specific properties, although the proposed model allows to take into consideration the biasness in the sample and to make inference for the population of interest.

Before using the ABC rejection algorithm, under both weigh functions $w_3$ and $w_4$, one should also adopt the prior distributions for the components of $\zeta$. Initially, the parameters of the prior distributions are selected is such a way to explore the parameter space in a large range of values. Additionally, any prior information should be incorporated in the selection of the parameters of the prior distributions. For that reasons the prior distributions of $\mu_x$ and $\mu_y$ were set to be the $N(\breve{\mu}_x, \breve{\sigma}_x)$ and $N(\breve{\mu}_y, \breve{\sigma}_y)$, where $\breve{\mu}_x$, $\breve{\mu}_y$, $\breve{\sigma}_x$ and $\breve{\sigma}_y$ denote the mean and the standard deviation of the observed $x_i s$ and $y_i s$, respectively. The prior distributions of $\sigma_x$, $\sigma_y$, $\gamma_1$ and $\gamma_2$ were set to be $TN(\breve{\sigma}_x, \breve{\sigma}_x/3)$, $TN(\breve{\sigma}_y, \breve{\sigma}_y/3)$, $TN(20, 20/3)$ and $TN(20, 20/3)$, respectively, where $TN$ denotes the truncated normal distribution at zero. For the parameter $\rho$ the linear transformation $2W - 1$ of the $W \sim beta(\alpha, \beta)$ distribution was used as prior distribution. The parameters of the transformed beta distribution were set equal to 3 and 5, respectively, in order for the mean and the mode to be negative ($-0.25$ and $-0.33$, respectively) and not to have a relative sharp prior. These values of the parameters $\alpha$ and $\beta$ and the corresponding negative mean and mode of the prior distribution, reflect our prior knowledge that the brain metabolism and so FDG PET score is decreasing by advancing age.

**TABLE 1** Descriptive statistics of the posterior distributions of the parameters of the bivariate normal distribution for the AD data under the incorrect $w_3$ (upper half) and the correct $w_4$ (lower half) weight functions based on 30,000 simulated samples

| Weight fun. | Descr. stat. | $\mu_x$ | $\sigma_x$ | $\mu_y$ | $\sigma_y$ | $\rho$ | $\gamma_1$ | $\gamma_2$ |
|---|---|---|---|---|---|---|---|---|
| $w_3$ | 2.5% | 67.2087 | 5.2172 | 1.0805 | 0.0677 | −0.6934 | 6.7571 | 6.1368 |
| Nonrejected | 25% | 71.0854 | 7.6561 | 1.1324 | 0.1014 | −0.3953 | 14.6434 | 14.5557 |
| samples: 16980 | 50% | 73.1296 | 9.1251 | 1.1607 | 0.1217 | −0.1949 | 18.7840 | 18.9306 |
| | 75% | 75.1932 | 10.6573 | 1.1900 | 0.1422 | 0.0248 | 22.9916 | 23.5078 |
| | 97.5% | 78.8922 | 13.5811 | 1.2429 | 0.1810 | 0.4277 | 30.9346 | 32.1594 |
| | mean | 73.1259 | 9.1978 | 1.1611 | 0.1223 | −0.1782 | 18.8218 | 19.0092 |
| | st.dev | 2.9962 | 2.1576 | 0.0420 | 0.0293 | 0.2936 | 6.1800 | 6.5923 |
| $w_4$ | 2.5% | 73.7070 | 5.2534 | 0.9621 | 0.0697 | −0.6167 | 5.9535 | 6.3947 |
| Nonrejected | 25% | 77.7724 | 7.7043 | 1.0164 | 0.1019 | −0.2484 | 14.9456 | 14.7380 |
| samples: 14901 | 50% | 80.0921 | 9.1946 | 1.0470 | 0.1208 | −0.0163 | 19.8402 | 19.1889 |
| | 75% | 82.4436 | 10.7253 | 1.0777 | 0.1408 | 0.2261 | 24.8157 | 23.6558 |
| | 97.5% | 86.8817 | 13.5772 | 1.1313 | 0.1789 | 0.6164 | 34.2791 | 32.1414 |
| | mean | 80.1438 | 9.2507 | 1.0468 | 0.1217 | −0.0099 | 19.9289 | 19.2324 |
| | st.dev | 3.3994 | 2.1676 | 0.0440 | 0.0282 | 0.3253 | 7.2159 | 6.5737 |

After the nonrejection of 50 samples, that is, after 50 retained observations of $\zeta^*$, the prior distribution of each component was updated using the information gained by these 50 observations. More specifically, the mean and the standard deviation of the respective retained values of the parameters $\mu_x, \mu_y, \sigma_x, \sigma_y$, $\gamma_1$ and $\gamma_2$ were used as the mean and the standard deviation of the corresponding normal prior distributions or as the $\mu_j$ and $c_j$ parameters of the corresponding truncated at zero normal prior distribution. The beta prior distribution for $\rho$ was updated to $beta(3, 3(1-\bar{r})/(\bar{r}+1))$, so that the mean value of the updated prior distribution to be equal with the mean value $\bar{r}$ of the retained values of $\rho$.

## 4.3 | Results and discussion

Table 1 presents the descriptive statistics of the corresponding p.d.f. of the posterior distributions of the parameters based on 16980 and 14901 (56.6% and 49.67%, respectively) nonrejected samples out of 30,000 simulated samples for both weight functions $w_3$ (upper half) and $w_4$ (lower half). From these results one can conclude that the median and the mean of the posterior distribution of $\rho$, which is the key parameter of the application, is clearly negative under both models. Additionally, under both models, approximately 75% of the accepted values for $\rho$ is smaller or slightly larger than zero. This result agrees with our prior knowledge that brain metabolism and so FDG PET is affected by advancing age. Therefore, it seems that both models provide similar results regarding the posterior distributions of $\rho$. A similar conclusion holds also for the posterior distributions of $\gamma_1$ and $\gamma_2$ and the standard deviations $\sigma_x$ and $\sigma_y$ of the random variables since the posterior distributions for these parameters present similar characteristics under the two models. It is worth mentioning that the posterior distributions of $\gamma_1$ and $\gamma_2$ seem to prefer relative large values. This may reflect the severe biasness induced to the sample by observing only patients with AD dementia.

Nevertheless, different behavior is revealed, when the posterior distributions of the means of the two variables are compared. For example, under the correct model $(w_4(x, y; , \gamma_1, \gamma_2))$ the mean age of the population of interest is estimated to be little over of 80 years, while under the incorrect model $(w_3(x, y; , \gamma_1, \gamma_2))$ is estimated just over 73 years, a great difference. A similar observation can be made and for the estimated mean of the FDG PET. Actually, the mean FDG PET score under the one model is higher than the mean FDG PET score under the other model. These remarks are actually consistent with the differences between the two models and can serve, combined with any prior knowledge on the population, as a confirmation of the correct choice between the two candidate weight functions. In the present application, as it is already stated, the mean age of the population of interest is more likely to be a value close to 80 and for sure larger than the mean age in the sample ($\bar{x} = 75.253$), since the observed sample is based on a cohort that does not reflect a naturalistic clinical setting, due to the advertising strategies that were employed for participant recruitment. For such studies, it has already been shown that the age of patients is indeed lower compared to that of nonuniversity centers (Weih et al., 2009).

Moreover, volunteers of such studies are characterized by high sensibility regarding cognitive impairment and seek advice and help earlier, when they are confronted by relatively mild memory deficits, in comparison to patients of pure clinical settings (Grimmer et al., 2015; Weih et al., 2009). Thus, less advanced hypometabolism and subsequently higher FDG PET scores are

expected in participants of initiatives such as ADNI than in other individuals with memory complaints. All, the above remarks are illustrated in Figure 2, in which the contour plots of the two estimated bivariate normal distributions under the incorrect model $w_3(x, y; , \gamma_1, \gamma_2)$ (left plot) and under the correct model $w_4(x, y; , \gamma_1, \gamma_2)$ (right plot) using the mean values for the parameters (denoted with solid lines) from the posterior distributions of the parameters are embedded in the scatterplot of the observed sample for the AD dementia data. It is clear, that both procedures managed to alter the false positive observed correlation to negative for the population. The plots also reveal the different characteristics of the two weight functions. In the left plot, it is assumed incorrectly that individuals with large age or/and small FDG PET are overrepresented in the observed sample, although such individuals are, in reality, underrepresented due to the nature of the sampling procedure.

*Remark* 4.3. Ignoring incorrectly the biasness in the sample and treating the sample as a random one, the maximum likelihood estimators (mle) for the bivariate normal model are: $\hat{\mu}_x = 75.2526, \hat{\mu}_y = 1.07613, \hat{\sigma}_x = 8.5317, \hat{\sigma}_y = 0.1203$ and $\hat{\rho} = 0.4762$. Comparing the mles and the descriptive statistics of the corresponding p.d.f. of the posterior distributions under the correct model, the following remarks can be made. First, the maximum likelihood underestimates the mean age and overestimates the mean FDG PET score of the population of interest, since $\hat{\mu}_x$ is close to the 2.5% percentile point of the posterior distribution of $\mu_x$ and $\hat{\mu}_y$ is larger than the 75% percentile point of the corresponding posterior distribution. Second, only a small portion ($< 2.5\%$) of the accepted values of the correlation coefficient $\rho$ is larger than the value of the mle ($\hat{\rho} = 0.4762$). Finally, the mles of the standard deviations $\sigma_x$ and $\sigma_y$ do not suffer from the same problems as the mles of the mean values.

Source code to reproduce the results of this section (Table 1 and Figure 2) as well as Figure 1 is available as Supporting Information on the journal's web page (http://onlinelibrary.wiley.com/doi/10.1002/bimj.201900046/suppinfo).

## 5 | DISCUSSION

The proposed modeling approach of Berkson's paradox relies on the fact that Berkson's paradox is actually a selection bias problem, which occurs when individuals satisfying specific properties are overrepresented in the sample. Therefore, the use of the weighted distributions is a natural choice for modeling such data. The weight functions manage to describe different scenarios and degrees of biasness and allow, using the ABC rejection algorithm, to make inference on the population of interest.

As with any Bayesian approach, the proposed ABC rejection algorithm is effected by the choice of the prior distributions. Prior distributions, in general, can have a significant effect on the posterior distributions especially in cases in which a relatively small number of data are available or the prior distributions are extremely sharp. In such cases, the posterior distributions are highly effected by the prior distribution which reflects the prior knowledge about the parameters. On the contrary, using noninformative priors results in posterior distributions that are mainly defined by the data. Since for Berkson's paradox some prior knowledge is expected to be available, at least for some parameters, the use of informative priors is recommended whenever is possible. This approach was illustrated in a real data application for AD dementia.

The proposed method can be applied to r.v.s with known, based on previous knowledge, parametric form. If such information is absented, model selection criteria using Bayes factors can be applied (for more details, see Marin, Pillai, Robert, & Rousseau, 2014). Finally, it is an open problem to generalize the method to deal with missing or censored data.

## ACKNOWLEDGMENTS

Authors would like to thank the two anonymous reviewers and the Associate Editor for critically reading the manuscript and suggesting substantial improvements. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche, Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO, Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer, Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute

for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

## CONFLICT OF INTERESTS

The authors have declared no conflict of interest.

## ORCID

*Polychronis Economou* https://orcid.org/0000-0001-6452-5920
*Apostolos Batsidis* https://orcid.org/0000-0002-4491-4387
*George Tzavelas* https://orcid.org/0000-0002-3284-4342
*Panagiotis Alexopoulos* https://orcid.org/0000-0002-5742-1733

## REFERENCES

Afonso, L., & Corte Real, P. (2016). Using weighted distributions to model operational risk. *ASTIN Bulletin*, *46*(2), 469–485.

Alavi, S., & Chinipardaz, R. (2007). The dependence structure in bivariate weighted distributions. *Far East Journal of Theoretical Statistics*, *23*(1), 19–29.

Alexopoulos, P., Grimmer, T., Perneczky, R., Domes, G., & Kurz, A. (2006). Progression to dementia in clinical subtypes of mild cognitive impairment. *Dementia and Geriatric Cognitive Disorders*, *22*, 27–34.

Alzheimer's Association (2016). 2016 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *12*(4), 459–509.

Arnold, B., & Nagaraja, H. (1991). On some properties of bivariate weighted distributions. *Communications in Statistics - Theory and Methods*, *20*(5-6), 1853–1860.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, *2*, 47–53.

Csilléry, K., Blum, M., Gaggiotti, O., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418.

Diggle, P. J., & Gratton, G. J. (1984). Monte Carlo statistical methods of inference for implicit statistical models. *JRSC- Series B*, *46*(2), 193–227.

Drew, L. (2018). An age-old story of dementia. *Nature*, *559*, S2–S3.

Duong, T., Goud, B., & Schauer, K. (2012). Closed-form density-based framework for automatic detection of cellular morphology changes. *Proceedings of the National Academy of Sciences*, *109*(22), 8382–8387.

Feinstein, A., Walter, S., & Horwitz, R. (1986). An analysis of Berkson's bias in case-control studies. *Journal of Chronic Diseases*, *39*(7), 495–504.

Fisher, R. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, *6*(1), 13–25.

Fritsch, T., McClendon, M., Wallendal, M., Hyde, T., & Larsen, J. (2014). Prevalence and cognitive bases of subjective memory complaints in older adults: Evidence from a community sample. *Journal of Neurodegenerative Diseases*, *2014*, 176843.

Geneletti, S., Best, N., Toledano, M. B., Elliot, P., & Richardson, S. (2013). Uncovering selection bias in case-control studies using Bayesian post-stratification. *Statistics in Medicine*, *32*, 2555–2570.

Greenland, S. (2003). Quantifying biases in casual models: Classical confounding vs. collider-stratification bias. *Epidemiology*, *14*, 300–306.

Grimmer, T., Beringer, S., Kehl, V., Alexopoulos, P., Busche, A., Förstl, H., … Kurz, A. (2015). Trends of patient referral to a memory clinic and towards earlier diagnosis from 1985–2009. *International Psychogeriatrics*, *27*(12), 1939–1944.

Gupta, R. C., & Kirmani, S. (1990). The role of weighted distributions in stochastic modeling. *Communications in Statistics: Theory and Methods*, *19*(9), 3147–3162.

Hernan, M., Hernandez-Diaz, S., & Robins, J. (2004). A structural approach to selection bias. *Epidemiology*, *15*, 615–625.

Ishibashi, K., Onishi, A., Fujiwara, Y., Oda, K., Ishiwata, K., & Ishii, K. (2018). Longitudinal effects of aging on $^{18}$F-FDG distribution in cognitively normal elderly individuals. *Scientific Reports*, *8*(1), 115579.

Izadkhah, S., Amini, M., & Borzadaran Mohtashami, G. R. (2016). Preservation of dependence concepts under bivariate weighted distributions. *Communications in Statistics - Theory and Methods*, *45*(15), 4589–4599.

Jack, C., Bennett, D., Blennow, K., Carrillo, M., Dunn, B., Haeberlein, S., … Sperling, R. (2018). NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, *14*(4), 535–562.

Jain, K., & Nanda, A. (1995). On multivariate weighted distributions. *Communications in Statistics-Theory and Method*, *24*(10), 2517–2519.

Jiang, J., Sun, Y., Zhou, H., Li, S., Huang, Z., Wu, P., … Alzheimer's Disease Neuroimaging Initiative (2018). Study of the influence of age in $^{18}$F-FDG pet images using a data-driven approach and its evaluation in Alzheimer's disease. *Contrast media & molecular imaging*, *2018*, 3786083.

Mahfoud, M., & Patil, G. (1982). On weighted distributions. In *Statistics and Probability: Essays in Honor of C.R. Rao* (pp. 479–492). North-Holland, New York: John Wiley & Sons, Ltd.

Marin, J., Pillai, N., Robert, C., & Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(5), 833–859.

Morgan, L. S. (Ed.). (2013). *Handbook of causual analysis for social research*. New York: Springer.

Nanda, A., & Jain, K. (1999). Some weighted distribution results on univariate and bivariate cases. *Journal of Statistical Planning and Inference*, *77*(2), 169–180.

Navarro, J., Ruiz, J., & Aguila, Y. D. (2006). Multivariate weighted distributions: A review and some extensions. *Statistics*, *40*(1), 51–64.

Nelsen, R. B. (2006). *An introduction to copulas*. New York: Springer-Verlag.

Patil, G., & Rao, C. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, *34*(2), 179–189.

Pearl, J. (1995). Casual diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Peritz, E. (1984). Berkson's bias revisited. *Journal of Chronic Diseases*, *37*(12), 909–916.

Prince, M., Knapp, M., Guerchet, M., McCrone, P., Prina, M., Comas-Herrera, A., … Salimkumar, D. (2014). *Dementia UK: Second Edition — Overview*. London, UK: Alzheimer's Society.

Rao, C. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhyā: The Indian Journal of Statistics, Series A*, *27*(2/4), 311–324.

Rasmussen, K. L., Tybjæaerg-Hansen, A., Nordestgaard, B. G., & Frikke-Schmidt, R. (2018). Absolute 10-year risk of dementia by age, sex and apoe genotype: A population-based cohort study. *Canadian Medical Association Journal*, *190*(35), E1033–E1041.

Rice, L., & Bisdas, S. (2017). The diagnostic value of FDG and amyloid pet in Alzheimer's disease: A systematic review. *European Journal of Radiology*, *94*, 16–24.

Robb, C., Udeh-Momoh, C., Wagenpfeil, S., Schope, J., Alexopoulos, P., & Perneczky, R. (2017). Biomarkers and functional decline in prodromal Alzheimer's disease. *Journal of Alzheimer's Disease*, *58*, 69–78.

Rotnitzky, A., & Robins, J. (2005). Inverse probability weighted estimation in survival analysis. In *Encyclopedia of Biostatistics*. New York: John Wiley & Sons, Ltd.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, *12*(4), 1151–1172.

Samuelsen, S., Anestad, H., & Skrondal, A. (2007). Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics*, *343*, 103–119.

Sarabia, J. M., & Gomez-Deniz, E. (2008). Construction of multivariate distributions: A review of some recent results. *Statistics and Operations Research Transactions*, *32*(1), 3–36.

Schroeder, D. (2010). *Accounting and causal effects: Econometric challenges*. New York: Springer.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Cambridge, MA: The MIT press.

VanderWeel, T., Herman, M., & Robins, J. (2008). Casual directed acyclic graphs and the direction of unmeasured confoundin bias. *Epidemiology*, *19*, 720–728.

Warda, A., Arrighib, H., Michelsa, S., & Cedarbaum, J. (2011). Mild cognitive impairment: Disparity of incidence and prevalence estimates. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *8*(1), 24–21.

Weih, M., Sulimma, A.-K., Lehfeld, H., Niklewski, G., Sonnenberg, M., Richter-Schmidinger, T., … Kornhube, J. (2009). Comparison of patient therapy adherence of two structural different memory clinics. *Fortschr Neurol Psychiatr*, *77*(4), 212–217.

Westreich, D. (2012). Berkson's bias, selection bias, and missing data. *Epidemiology*, *23*(1), 159–164.

## SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

**How to cite this article:** Economou P, Batsidis A, Tzavelas G, Alexopoulos P, Alzheimer's Disease Neuroimaging Initiative. Berkson's paradox and weighted distributions: An application to Alzheimer's disease. *Biometrical Journal*. 2020;62:238–249. https://doi.org/10.1002/bimj.201900046