Appendix

Numerical example of Berkson's fallacy

Here we provide a numerical example of Berkson's fallacy. Table 1 below shows one of the theoretical lay-out and probabilities from one of the original examples worked out by Berkson.[1] The study assesses the association between cholecystic disease and diabetes mellitus. Controls are subjects with ophthalmologic refractive errors. The example is constructed so that the three diseases are independent in the population.

With a source population of 10.000.000 subjects, the following two-by-two tables (Table 2a-2c) can be constructed based on the numbers in Table 1, assuming independent hospitalization probabilities. Furthermore, it is assumed that subjects with both the case and control disease are counted (only) as cases. Berkson uses as measure of association the case-control difference in the "incidence" of D1 (actually these were prevalences), but the particular choice of association measure is irrelevant for our exposition (we use odds ratios in our worked out examples, following Boyd).For details about the calculations we refer to the papers of Berkson and Boyd.[1,4] It follows that the odds ratio for the association between cholecystic disease and diabetes is 1 in the total population (by definition), 1.89 among hospitalized subjects and again 1 in subjects out of the hospital.

Magnitude and direction of Berkson's fallacy

The example above can be extended to general formulas to calculate the relative magnitude of Berkson's fallacy. The formula's have been derived by Boyd.[4]

Let p_1, p_2 and p_3 denote the prevalence of diseases D1, D2, and D3, respectively. In Berkson's scenario p_1, p_2 and p_3 are independent and thus the D1-D2 odds ratio in the source population is 1. Let h_1, h_2 and h_3 denote the probability of hospitalization of diseases D1, D2, and D3, respectively. It is assumed that the hospitalization probabilities are also independent. When subjects with both the case and control disease are counted as cases, as Berkson did, the D1-D2 odds ratio indicating the magnitude of the fallacy in the hospitalized population equals:

$$OR = \frac{\left[1-(1-h_1)(1-h_2)(1-p_3h_3)\right]h_3}{\left[h_2+(1-h_2)p_3h_3\right](h_3+h_1-h_3h_1)} \qquad \text{(Equation 1)}$$

When subjects who have both the case and control disease are excluded, the magnitude of the fallacy depends only on the three hospitalization probabilities:

$$OR = \frac{\left[h_2+h_1(1-h_2)\right]h_3}{\left[h_3+h_1(1-h_3)\right]h_2} \qquad \text{(Equation 2)}$$

These equations lead to some additional observations:[4,10]

1) Given a low prevalence of the control disease (in case of overlap between cases and controls, equation 1), Berkson's fallacy is negative when $h_2 > h_3$, and positive when $h_3 > h_2$.

2) The lower h_1, the smaller the bias will be. When the exposure does not lead to hospitalization at all (h_1=0), these formulas simplify to 1. In terms of graph theory, hospitalization is no longer a collider of exposure and outcome because there is no arrow from D1 to H.

Those relationships have been illustrated with some numeric examples in Table 3. The first row represents Berkson's original example also mentioned above. When the odds ratio in the source population does not equal 1, the in-hospital odds ratio is the product of the population odds ratio and odds ratio due to Berkson's fallacy following the above formulas.

These formulas apply only to Berkson's setting in which diseases lead to hospitalization through independent mechanisms. In all other scenarios, e.g. clinical decisions to hospitalize because of the presence of multiple diseases, there is additional bias, beyond the fallacy as quantified in the formulas (see text).

Magnitude of the indirect Berkson's fallacy

Similar formulas and rules-of-thumb apply to indirect Berkson's fallacy in exposure-disease associations as to the classical Berkson's fallacy in disease-disease associations, although in the equations all hospitalization probabilities of D1 (h_1) should be changed into the hospitalization probability of subjects with the exposure E (via D1). As not all exposed subjects will have this other disease D1, the magnitude of indirect Berkson's fallacy in will be generally lower than the classical Berkson's fallacy.

Appendix Tables

Table 1: Original example of Berkson's fallacy: prevalence and hospitalization probabilities

		Population prevalence		Hospitalization probability		Working example
Disease 1	Exposure	p1	**0.03**	h1	**0.15**	(Cholecystic disease)
Disease 2	Cases	p2	**0.01**	h2	**0.05**	(Diabetes mellitus)
Disease 3	Controls	p3	**0.1**	h3	**0.2**	(Refractive errors)

In Berkson's example, Exposure (D1) was cholecystic disease; Case (D2) was diabetes, and Control (D3) was ophthalmologic refractive disorders.

Table 2a: Association in the overall population

	Exposed	**Unexposed**	
Cases	3.000	97.000	100.000
Controls	29.700	960.300	990.000
	32.700	1.057.300	1.090.000

Association with cholecystic disease in the overall population.; the odds ratio is 1.

Table 2b: Association in the hospitalized population

	Exposed	**Unexposed**	
Cases	626	6.693	7.319
Controls	9.504	192.060	201.564
	10.130	198.753	208.883

Association with cholecystic disease in the hospitalized population; the odds ratio is 1.89.

Table 2c: Association in the non-hospitalized population

	Exposed	**Unexposed**	
Cases	2.374	90.307	92.681
Controls	20.196	768.240	788.436
	22.570	858.547	881.117

Association with cholecystic disease in the non-hospitalized population. The odds ratio is 1.

Table 3: Examples of the magnitude of classical Berkson's fallacy with varying hospitalization probabilities

h_1	h_2	h_3	p_3	OR_1	OR_2
0.15	0.05	0.2	0.1	1.89	2.41
0.01	0.05	0.2	0.1	1.09	1.14
0	0.05	0.2	0.1	1.00	1.00
0.15	0.2	0.05	0.1	0.41	0.42
0.01	0.2	0.05	0.1	0.87	0.87
0	0.2	0.05	0.1	1.00	1.00
0.15	0.2	0.2	0.1	0.97	1.00
0.15	0.05	0.05	0.1	0.93	1.00

h_1, h_2 and h_3 are the hospitalization probabilities of the "exposure" disease (D1), the case disease (D2) and the control disease (D3), which are independent. All calculations assume a prevalence of the control disease of 0.1. The odds ratios (OR) indicate the magnitude of the spurious association that arises by Berkson's fallacy, assuming a causal OR of 1. The first OR represents the fallacy in the situation in which patients who have both the case and the control disease are counted (only) as cases (Equation 1); the second OR follows when those patients with both diseases are excluded (Equation 2). The first row shows Berkson's original example.