Chapter 4
Four Approaches to the Reference Class Problem

Christian Wallmann and Jon Williamson

Abstract We present and analyse four approaches to the reference class problem. First, we present a new objective Bayesian solution to the reference class problem. Second, we review Pollock's combinatorial approach to the reference class problem. Third, we discuss a machine learning approach that is based on considering reference classes of individuals that are similar to the individual of interest. Fourth, we show how evidence of mechanisms, when combined with the objective Bayesian approach, can help to solve the reference class problem. We argue that this last approach is the most promising, and we note some positive aspects of the similarity approach.

4.1 Introduction

The problem of determining the probability that a particular patient has a certain attribute (e.g., has a certain disease, or has a certain prospect of recovery) is of fundamental importance to medical diagnosis, prediction and treatment decisions. Theories of *direct inference* aim to solve this problem of the single-case. In direct inference, single-case probabilities are often calibrated to estimates of chances in reference classes to which the individual of interest belongs. The major problem in direct inference is to determine an appropriate single-case probability when an individual belongs to several reference classes for which data is available, and where estimates of chances differ from reference class to reference class. This is the *reference class problem*.

Let us consider how direct inference sometimes proceeds by means of an example. The question is whether Nataly, a patient with breast cancer, will survive at least five more years. The doctor may know the size of the index lesion (S), node status (N) and (G) grade of tumour of the patient. She then calculates the Nottingham prognostic index (NPI) score by NPI = $[0.2 \times \mathrm{S}] + \mathrm{N} + \mathrm{G}$ (Haybittle et al. 1982). Let's suppose Nataly has an NPI-score of 4.2. There is statistical

C. Wallmann (✉) • J. Williamson
Department of Philosophy, University of Kent, Canterbury, UK
e-mail: c.wallmann-520@kent.ac.uk

G. Hofer-Szabó, L. Wroński (eds.), *Making it Formally Explicit*, European Studies in Philosophy of Science 6, DOI 10.1007/978-3-319-55486-0_4

evidence that 7 out of 10 people with an NPI-score between 3.4 and 5.4 survive for more than 5 years. One might infer, then, that Nataly has a probability of survival of 70%.

As straightforward as this may seem, in practice it is rather difficult to determine appropriate reference classes. Often an individual belongs to many populations for which there is statistical information about the chance of the attribute of interest in those populations. In our example, the patient will have a certain nationality, a certain attitude towards religion, a certain social status, and a certain genetic profile, and there may be evidence of chances available for several of these reference classes. While we might assume that religious belief is irrelevant for survival time, this is less clear for nationality and social status, and certainly false for genetic profile. The most intractable problem for direct inference is the problem of conflicting reference classes. The problem occurs when chances are available for two (or more) reference classes to which the individual of interest belongs, these chances differ, and there is no chance available for the intersection of these reference classes. John Venn remarked with respect to John Smith, an Englishman that has tuberculosis,

> Let us assume, for example, that nine out of ten Englishmen are injured by residence in Madeira, but that nine out of ten consumptive persons are benefited by such a residence. These statistics, though fanciful, are conceivable and perfectly compatible. John Smith is a consumptive Englishman; are we to recommend a visit to Madeira in his case or not? In other words, what inferences are we to draw about the probability of his death? Both of the statistical tables apply to his case, but they would lead us directly contradictory conclusions. [...] Without further data, therefore we can come to no decision. (Venn 1888, p. 222–223)

Suppose that we know that an individual c belongs to two reference classes B and C, written Bc and Cc. Suppose further that we know the chance of the target attribute A in the reference class B as well as in the reference class C, i.e., $P^*(A|B) = r$ and $P^*(A|C) = s$ (and nothing else). The problem of conflicting reference classes is the problem of determining the probability $P(Ac)$ that c has attribute A.

A word on notation. We use P^* to represent the objective chance distribution. This is *generic* in the sense that it is defined over attributes, classes and variables which are repeatedly instantiatable. We make no metaphysical assumptions about chance here: a chance might be understood as a dispositional attribute (or propensity), a long-run frequency, an objectivised subjective probability, or posited by a Humean account of laws, for example. We use *freq* to denote the sample frequency. Again, this is generic. It is the probability distribution induced by a sample or a dataset, and it may be used to estimate the chance distribution P^*, i.e., the data-generating distribution. Finally, we take other probability functions, such as P, $P^\dagger$, to be *single-case* probability functions, to be used for direct inference. These are single-case in the sense that they are defined over propositions or events which are not repeatedly instantiable.

In Sect. 4.2, we develop a new objective Bayesian solution to the problem of conflicting reference classes. In Sect. 4.3, we review Pollock's approach to the problem. We show that it is based on mistaken assumptions. In Sect. 4.4, we relate similarity-based machine learning techniques to the reference class problem. All these approaches are classifiable as generic-probability approaches: chances

in reference classes are estimated by frequencies induced by datasets and those estimates are then aggregated to obtain a probability for the single case. In Sect. 4.5, we briefly discuss two key challenges that face generic-probability approaches: the problem of small sample size, which arises when reference classes are so narrowly defined that it is difficult to obtain samples to estimate the corresponding chances, and the problem of inconsistent marginals, which arises when different samples yield incompatible sample frequencies. We show that the objective Bayesian approach can address these challenges by appealing to confidence region methods, and that the similarity approach can be re-interpreted as a single-case-probability approach in order to avoid these difficulties. In Sect. 4.6, we show how to make use of evidence of mechanisms to help solve the reference class problem. Evidence of mechanisms helps in two ways: it can help by enriching the structure of the problem formulation, and it can help to extrapolate evidence of chances from reference classes that are not instantiated by the particular individual of interest to those that are. Either way leads to more credible direct inferences. We conclude by discussing the circumstances in which the various methods are appropriate in Sect. 4.7.

4.2 An Objective Bayesian Approach

According to the version of objective Bayesian epistemology developed by Williamson (2010), one can interpret predictive probabilities as rational degrees of belief, and these rational degrees of belief are obtained by taking a probability function, from all those that satisfy constraints imposed by evidence, that has maximal entropy.[1] That degrees of belief should be probabilities is called the Probability norm. That they should satisfy constraints imposed by evidence is the Calibration norm. In particular, degrees of belief should be calibrated to chances, insofar as the evidence determines these probabilities. That a belief function should have maximal entropy is the Equivocation norm. The maximum entropy function is interpretable as the most equivocal or least committal probability function (Jaynes 1957). The Equivocation norm is justifiable on the grounds of caution: a maximum entropy function is a probability function which minimises worst-case expected loss (Williamson 2017a, Chapter 9).

Williamson (2013) locates the reference class problem at the stage of the Calibration norm:

> The infamous reference class problem must be tackled at this stage, i.e., one must decide which items of evidence about the generic physical probabilities should be considered when determining single case probabilities [...]. (Williamson 2013, p. 299)

[1] The entropy of a probability function P defined on a set of logically independent propositions $\{E_1, \ldots, E_n\}$ is defined by $-\sum_{i=1}^{n} P(E_i) \log P(E_i)$.

To solve the problem of conflicting reference classes, Williamson draws on other approaches to direct inference. Without endorsing it, he discusses Kyburg's approach as one possible option. A combination of Williamson's and Kyburg's approach first combines the information in conflicting reference classes to obtain an interval for $P(Ac)$ and then applies the Equivocation norm to this already aggregated degree of belief. $\{P^*(A|B) = 0.9, P^*(A|C) = 0.4, Bc, Cc\}$, for instance, yields by Calibration $P(Ac) \in [0.4, 0.9]$ and subsequently by Equivocation $P(Ac) = 0.5$. However, as we are going to see now, the objective Bayesian approach has sufficient resources to solve the problem of conflicting reference classes on its own. Rather than considering it purely as a calibration problem, the proposal presented here spreads the load between the Calibration norm and the Equivocation norm.

The objective Bayesian approach presented here proceeds first by calibrating, then by aggregating. First, it calibrates conditional degrees of belief to estimates of chances, $P(Ac|Bc) = P^*(A|B) = r$ and $P(Ac|Cc) = P^*(A|C) = s$. Second, it equates the direct inference probability that c belongs to the target class A with $P^\dagger(Ac|Bc \wedge Cc)$, where $P^\dagger$ satisfies the constraints $P^\dagger(Ac|Bc) = r$ and $P^\dagger(Ac|Cc) = s$ but is otherwise as equivocal as possible. We then have

New objective Bayesian solution to the problem of conflicting reference classes. $P(Ac) = P^\dagger(Ac|Bc \wedge Cc)$, where $P^\dagger$ is a probability function that has maximal entropy among all probability functions P that satisfy $P(Ac|Bc) = r$ and $P(Ac|Cc) = s$.

The objective Bayesian approach combines the maximum entropy principle with probabilistic logic. The problem of determining the direct inference probability can be solved by linear programming and optimization techniques. If $P^*(A|B) = r$ and $P^*(A|C) = s$, the solution to the reference class problem is given by $P^\dagger(Ac|Bc \wedge Cc) = \frac{x_1}{x_1+x_5}$ where the vector x is the solution to the following optimization problem[2]:

$$\text{Maximise } -\sum_{i=1}^{8} x_i \log x_i \text{ subject to } Sx = b \text{ and } x_i \geq 0, \text{ where } b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \text{ and } S = \begin{pmatrix} r-1 & r-1 & 0 & 0 & r & r & 0 & 0 \\ s-1 & 0 & s-1 & 0 & s & 0 & s & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Although it can be difficult to provide analytic solutions to such problems, they can be solved numerically, by using, for instance, MAPLE or Matlab software.

To give an example, $\{P^*(A|B) = 0.9, P^*(A|C) = 0.4, Bc, Cc\}$ leads to $P(Ac) = 0.83$. Note that the objective Bayesian approach does not necessarily assign a very equivocal degree of belief to the proposition Ac. Indeed, it can lead to degrees of belief that are more extreme than either of the reference class frequencies. For instance, in absence of further constraints, $P^*(A|B) = 0.9, P^*(A|C) = 0.9$ leads to $P(Ac) = 0.96$. The reason for this is that the objective Bayesian approach leads

[2]See Wallmann and Kleiter (2014a,b) for a general procedure for generating the relevant optimization problem.

to beliefs that are equivocal *on average*. It may assign extreme degrees of belief to certain propositions and instead assign less extreme degrees of belief to other propositions in order to maximise the extent to which the belief function as a whole is equivocal.

In the next section, we will consider an approach that claims that more can be done to constrain $P^*(A|B \wedge C)$ before calibrating.

4.3 Pollock's Approach

Pollock's approach to direct inference involves first aggregating, then calibrating. Pollock first aggregates the values of the conflicting reference classes and estimates the value of $P^*(A|B \wedge C)$, and then calibrates $P(Ac)$ to the result. Since in Pollock's theory $P^*(A|B \wedge C)$ can be very well estimated (with probability 1), there is no role for equivocation to play.

Pollock motivates his theory of direct inference this way:

> Suppose we have a set of 10,000,000 objects. I announce that I am going to select a subset, and ask you how many members it will have. Most people will protest that there is no way to answer this question. It could have any number of members from 0 to 10,000,000. However, if you answer, "Approximately 5,000,000", you will almost certainly be right. This is because, although there are subsets of all sizes from 0 to 10,000,000, there are many more subsets whose sizes are approximately 5,000,000 than there are of any other size. In fact, 99% of the subsets have cardinalities differing from 5,000,000 by less than .08%. (Pollock 2011, p. 329)

This "peaking" property holds for finite sets and follows from elementary combinatorics. Moreover, the distribution of the subsets gets needle-like in the limit: the larger the set is, the greater the proportion of subsets that have size close to half the size of the set. Pollock takes this fact as a starting point for his theory of nomic probability. Pollock calls his theory 'nomic' because he is concerned with probabilities that are involved in statistical laws. Rather than being concerned with frequencies and relative frequencies involving actual events he is concerned with frequencies and relative frequencies among physically possible worlds. Probabilities therefore contain an irreducible modal element.

As a natural generalization from finite sets to infinite sets, Pollock's theory of nomic probabilities assumes that these peaking properties hold for infinite sets with probability 1.[3] If a peaking probability holds with probability 1, then the value

[3]Let pe be a point. It is called peaking point with probability 1 iff for all $\delta > 0$, $PROB(|P^*(A|S) - pe| < \delta) = 1$, i.e., for all $\epsilon > 0$, $PROB(|P^*(A|S) - pe| < \delta) > 1 - \epsilon$. If we think of δ being very small, for instance, 0.000001, then this means that almost all subsets S are such that $pe - 0.000001 \leq P^*(A|S) \leq pe + 0.000001$. Note that probability 1 does not mean that there are no exceptions, i.e., even if $PROB(|P^*(A|S) - pe| < \delta) = 1$, there are S such that $|P^*(A|S) - pe| > \delta$. However, such S are comparably few.

around which the nomic probabilities peak is called the *expectable value*. This leads to the following default independence principle:

If $P^*(A|B) = x$, then the expectable value for $P^*(A|B \wedge C)$ is x.

Pollock's solution to the problem of conflicting reference classes is to set the direct inference probability to the expectable value of the target attribute in the intersection of the conflicting reference classes (Pollock 2011). Moreover, he calculates the relevant expectable value. Let $P^*(A|B)$ denote the nomic probability of A given B (Pollock 2011). If, in addition to the reference class probabilities $P^*(A|B) = r$ and $P^*(A|C) = s$, the "base rate" $P^*(A|U) = a$ is given where $B, C \subseteq U$, Pollock shows that the expectable value of $P^*(A|B \wedge C)$ exists and is given by

$$Y(r, s|a) = \frac{rs(1-a)}{a(1-r-s)+rs} \ .$$

The Y-function can be used to tackle the problem of conflicting reference classes. If there is no knowledge of the chance of the target attribute A in some joint upper class U of B and C—i.e., if for no such U $P^*(A|U)$ is available—its expectable value p_0 can still be determined (Pollock 2011). Since the expectable value is attained with probability 1, the degree of under-determination for $P^*(A|B \wedge C)$ is very small and ignoring the Equivocation desideratum seems to be reasonable, i.e., we may equate $P(Ac)$ with the expectable value for $P^*(A|B \wedge C)$.

Pollock's solution to the problem of conflicting reference classes. The direct inference probability is given by

$$P(Ac) = Y(r, s|p_0) \ .$$

The expectable value p_0 of $P^*(A|U)$ is given by the first component of the solution (a, b, c) to the following system of equations (Pollock 2011).

$$\left(\frac{1-r}{1+(r-p_0)b}\right)^{1-r}\left(\frac{r}{-br+p_0}\right)^{r} = 1$$

$$\left(\frac{1-s}{1+(s-p_0)c}\right)^{1-s}\left(\frac{s}{-cs+p_0}\right)^{s} = 1$$

$$2p_0^3 - (-2br-2cs+b+c-3)p_0^2 + (2bcrs-bcr-bcs+bc+2br+2cs-b-c+1)p_0 - cbrs = 0$$

Although it is often difficult to provide explicit formulae for expectable values, they can be calculated numerically, by using Pollock's LISP-code (Pollock 2016).

Pollock's expectable value is relative to the probability distribution that is used to calculate the expectable value. The quality of the expectable value depends on the accuracy of this probability distribution. For his probability distribution,

Pollock extrapolates simple combinatorial probabilities from the finite to the infinite case. Doubt has been raised whether combinatorial probabilities are accurate in the domain of reference class reasoning (Wallmann 2017). Especially, the fact that $P^*(A|B \wedge C)$ is almost certainly very close to $Y(r, s|p_0)$ is difficult to reconcile with experience. In practice, we often find subsets that have a rather different chance of the target attribute than the original set. For instance, smoking rates in the United States vary strongly with gender, age, education, poverty status and many more. But according to Pollock's nomic probabilities, such variations seem to be almost impossible.

The mistake is this: the combinatorial probabilities in the finite case consider *arbitrary* subsets of the sets B and C. Every subset has the same relevance for direct inference. However, in the context of direct inference this is unreasonable. If we use a certain subset of B in practice, most likely we will not use a different but almost identical subset for direct inference. For instance, if we use the set of all Austrians, most likely we will not use the set of all Austrians except for Alexander for direct inference. Being Austrian but being not identical with Alexander is not expressing any causally relevant attribute. Thus, not every subset of a reference class is itself a reference class. We tend to consider classes of individuals that instantiate natural attributes—attributes which are causally relevant to the attribute of interest. Now, causally relevant attributes tend to be difference makers, i.e., they tend to be probabilistically dependent. Therefore, real reference-class probabilities vary to a greater extent than arbitrary-subset probabilities. Instead of clustering very closely around the expectable value, frequencies within sub-reference classes are more likely to cluster around multiple different values. Although expected values exist in the case of sub-reference classes, expectable values do not. While peaking properties hold with probability 1 for arbitrary subclasses, this fact is irrelevant to direct inference. To take another example, the proportion of smokers varies to a great extent between sub-reference classes of the reference class of all people living in the United States. The fact that most subsets of all people living in the United States share the same proportion of smokers with all people living in the United States is irrelevant to direct inference, because in direct inference we are only concerned with natural attributes as, for instance, gender, age, educational level. Peaking properties do not hold with probability 1 for classes that correspond to such natural attributes.

4.4 The Similarity Approach

Suppose that a reasonable measure of similarity between two individuals is available. The basic idea behind the similarity approach to direct inference is simple: we may predict whether an individual c will get a certain disease by considering the chance of disease in reference classes of individuals similar to the individual c. The more similar a reference class of individuals is to c, the more relevant information about the chance of disease in this class is for predicting whether c will get the disease.

An attribute-based similarity measure is based on the number of shared attributes. One way to define a attribute-based similarity measure is Gower's similarity coefficient (Gower 1971).[4] This measures the number of shared attributes of individuals in a reference class R and an individual c. Suppose that $R_1, \ldots, R_n$ are reference classes. Let $C_1, \ldots, C_N$ be the attributes that c is known to have (the C_i's may also contain negations). Let $C_j(R)$ denote the fact that all individuals in the reference class R have the attribute C_j.

$$sim_{Gow}(R, c) = \frac{|\{j \in \{1, \ldots, N\} : C_j(R)\}|}{N} \tag{4.1}$$

The attribute similarity solution to the reference class problem is given by:

Attribute similarity solution to the reference class problem. For $i = 1, \ldots, n$, let $P^*(A|R_i) = x_i$. Then the direct inference probability that c has disease A is given by

$$P(Ac) = T \sum_{i=1}^{n} sim_{Gow}(R_i, c) x_i, \tag{4.2}$$

where T is a normalizing constant. For instance, if Bc, Cc, Hc, Ec and $P^*(D|\neg E \wedge C) = x, P^*(D|B \wedge \neg C \wedge E \wedge H) = y, P^*(D|B \wedge C) = z$, then $P(Dc) = \frac{4}{6}(\frac{1}{4}x + \frac{3}{4}y + \frac{2}{4}z)$.

Observe that the direct inference probability may be influenced by chances in reference classes to which the individual does not belong. Hence, the similarity approach aims to solve a problem even more general than the problem of conflicting reference classes—the problem of applying chances in arbitrary classes to specific individuals. We shall return to this problem when we discuss extrapolation in Sect. 4.6.2.

Attribute-based similarity measures are commonly used in machine learning. Indeed, the approach advocated here is a special case of the machine learning technique called k-nearest neighbour weighted mean imputation; for details see Jerez et al. (2010, pp. 110–111). To impute a missing data value, a weighted average value of the k most similar reference classes is taken. Here, $k = n$, i.e., all reference classes are similar enough to contribute to the weighted average. More sophisticated similarity measures incorporate, for instance, base rates of diseases in the general population (Davis et al. 2010). However, attribute-based similarity measures do not distinguish between causally relevant and irrelevant attributes; every attribute is equally important.

[4]For a related but more sophisticated similarity measure see Davis et al. (2010).

4.5 Generic-Probability vs Single-Case-Probability Approaches

The objective Bayesian and Pollock's approach are generic-probability approaches. They follow the following procedure:

1. Group individuals to classes (reference classes).
2. Estimate the chance of the attribute of interest in each class (a generic probability).
3. Determine a direct inference probability from these values by maximum entropy or other techniques.

Generic-probability approaches face two fundamental challenges. One key obstacle is what we call the *problem of small sample size*. On the one hand, we would prefer *narrow* reference classes, i.e., classes of individuals which share many of the features instantiated by the individual in question, because such classes are particularly relevant to the individual. On the other hand, however, a narrow reference class is likely to contain few sampled individuals. Where this sample size is small, the sample frequency is likely to be rather different from the true chance that it is supposed to estimate. Thus a narrow reference class tends to yield an inaccurate estimate of the chance of the attribute of interest within the reference class. More generally, to arrive at a precise estimate of the data-generating chance distribution defined over many variables, a very large number of observations is needed, because a relatively small proportion of sampled individuals will share any particular combination of values of measured variables. A dataset measuring a large number of variables will often be too small to provide a reasonable estimate of the data-generating chance function.

A second key obstacle is what we call the problem of *inconsistent marginal probabilities*. This occurs when several samples are collected—several datasets are obtained—and certain variables occur with different frequencies in different datasets: there is then no joint probability function whose marginal probabilities match all the distributions determined by the datasets. This problem is very common. It may be attributable to bias, chance or to small sample sizes.

The following example, pitched at the objective Bayesian solution to the reference class problem, illustrates the two challenges.

- Suppose datasets D_1, D_2, D_3 measure sets of variables $V_1 = \{X_1, X_2, X_3\}, V_2 = \{X_2, X_3, X_4\}, V_3 = \{X_1, X_2, X_4\}$ respectively.
- The objective Bayesian approach seeks to find an appropriate joint probability function $P(X_1, X_2, X_3, X_4)$, by the following procedure:

  1. Marginals of the data-generating chance distribution P^* are estimated by the sample distributions determined by the datasets:

     – $P^*(X_1, X_2, X_3)$ is estimated by the observed frequency $freq_1(X_1, X_2, X_3)$ of D_1.

- $P^*(X_2, X_3, X_4)$ is estimated by the observed frequency $freq_2(X_2, X_3, X_4)$ of D_2.
- $P^*(X_1, X_2, X_4)$ is estimated by the observed frequency $freq_3(X_1, X_2, X_4)$ of D_3.

2. Consider the set $\mathbb{E}$ of all probability functions P that satisfy $P(X_1, X_2, X_3) = P^*(X_1, X_2, X_3)$, $P(X_2, X_3, X_4) = P^*(X_2, X_3, X_4)$, and $P(X_1, X_2, X_4) = P^*(X_1, X_2, X_4)$.
3. Determine the probability distribution $P^\dagger$ in $\mathbb{E}$ with maximum entropy.

The problem of small sample size arises in Step 1 if the datasets have too few observations to yield plausible estimates of the chances. The problem of conflicting marginals arises in Step 2. The dataset distributions $freq_1(X_1, X_2, X_3)$, $freq_3(X_1, X_2, X_4)$ determine the respective marginal distributions $freq_1(X_1, X_2)$, $freq_3(X_1, X_2)$. Typically, $freq_1(X_1, X_2) \neq freq_3(X_1, X_2)$, i.e., we have inconsistent marginals. Consequently, there is no probability function P that satisfies the above constraints.

The objective Bayesian approach has a potential line of response to these two challenges—a response that involves an appeal to *confidence regions* (Williamson 2017b, §4). As discussed above, the frequency distribution $freq_i$ determined by dataset D_i can be thought of as a point estimate of the marginal data-generating chance distribution $P^*(V_i)$, defined over the set V_i of variables measured by that dataset. For instance, $freq_1(X_1, X_2, X_3)$, the observed frequency distribution of dataset D_1, is treated as a point estimate of the data-generating chance distribution $P^*(X_1, X_2, X_3)$ in the above example. Now, a point estimate is almost always wrong. Rather than use a point estimate, one can instead infer that the data-generating chance distribution lies in a region around the point estimate—the confidence region R_i. This confidence region depends on the confidence level, i.e., how probable it is that similar samples will yield a sample distribution such that the chance distribution is contained in its confidence region. Thus the region corresponding to a 99% confidence level will be larger than that corresponding to a 95% confidence level: a larger confidence region is needed to increase the probability that the region contains the chance distribution. The confidence region method leads to a more subtle implementation of the Calibration norm: instead of calibrating degrees of belief to each dataset distribution (which is impossible when these marginal distributions conflict) one only needs to ensure that the belief function lies within the confidence region determined by each dataset (which is possible if one chooses a confidence level high enough to ensure that the regions do not conflict). Thus, instead of taking $\mathbb{E} = \{P : P(V_i) = freq_i(V_i)\}$, we take $\mathbb{E} = \{P : P(V_i) \in R_i(V_i)\}$. So in our example we have that $\mathbb{E} = \{P : P(X_1, X_2, X_3) \in R_1(X_1, X_2, X_3), P(X_2, X_3, X_4) \in R_2(X_2, X_3, X_4), P(X_1, X_2, X_4) \in R_3(X_1, X_2, X_4)\}$.

The confidence region approach also addresses the problem of small sample size. This is because the size of the confidence region depends on the number of individuals observed in the dataset and on the number of variables measured, as well as on the confidence level: fewer sampled individuals (or more measured variables) will lead to a wider confidence region and a less precise estimate of

the data-generating distribution, ceteris paribus. Moreover, wider regions typically correspond to a more equivocal maximum entropy probability function selected by the Equivocation norm. By choosing a confidence level that is both high enough that the confidence regions can be taken to be plausible estimates of the marginal chance distributions and high enough that the confidence regions do not conflict, one simultaneously solves the problem of inconsistent marginal probabilities and specifies a probability function which is somewhat influenced by the dataset distributions, but not unduly so when the sample size is small.

Having considered an objective Bayesian response to the two challenges, let us consider the other two approaches that we have encountered so far—Pollock's approach and the similarity approach.

Pollock does not address the two challenges. This must be considered to be a further point against Pollock's approach.

We presented the similarity approach as a generic-probability approach. Although it is not subject to the problem of inconsistent marginals, the problem of small sample size still remains for the generic version of the similarity approach. We will now see that the similarity approach can be reconstructed as purely single-case, in order to circumvent the problem.

Suppose that raw data from different studies that investigate a certain disease is available. Certain variables measured by one of these studies may not be measured by another study. For example, one study on acute kidney disease may measure clinical variables (age, gender, co-morbidities etc.), another may measure pathology variables (creatine testing, proteinuria testing etc.) and a third may measure variables from imaging procedures. In addition, in each study, parts of the study results may be missing for some participants.

We can represent the situation as one of missing data. To avoid unnecessary complications, we focus on binary variables. We consider the set of all individuals who have participated in at least one study. For each of the participants the data will consist of an entry for all of the N variables that are at least measured in one study. The entries are either NM (not measured), if the person did not participate in a study where the variable has been measured or if the person did participate but no value has been recorded, 1 if the participant has the attribute expressed by the variable and 0 if the person did participate in the study but does not have the attribute expressed by the variable. Formally, the data consists of observations for n-individuals, $b_1, \ldots, b_n$. For each individual b_i, $i = 1, \ldots, n$, the data consists of a string of information $Dat_i = (x_{i,1} \ldots x_{i,N})$; where for all $j = 1 \ldots N$, $x_{i,j} \in \{0, 1, NM\}$ specifies the status of the individual b_i with respect to the attribute X_j. Suppose that for each individual b_i, $i = 1, \ldots, n$, measurements on A are available, i.e., $A(b_i) \in \{0, 1\}$.

Here is an example for 6 patients $b_1, \ldots, b_6$ and four variables in three datasets $V_1 = \{X_1, X_2, X_3\}, V_2 = \{X_2, X_3, X_4\}, V_3 = \{X_1, X_2, X_4\}$:

Patient	X_1	X_2	X_3	X_4
b_1	1	1	1	*NM*
b_2	0	1	0	*NM*
b_3	*NM*	1	1	0
b_4	*NM*	0	0	1
b_5	1	0	*NM*	1
b_6	0	0	*NM*	1

Gower's similarity measure for individuals, rather than for attributes, measures the number of shared attributes of b_k and b_l divided by the number of attributes where data for both b_k and b_l are available:

$$sim_{Gow}(b_k, b_l) = \frac{|\{j \in \{1, \ldots, N\} : X_j(b_k) = X_j(b_l)|}{|\{j \in \{1, \ldots, N\} : X_j(b_k) \neq NM \wedge X_j(b_l) \neq NM\}|} \tag{4.3}$$

The direct inference probability for Ac is a weighted average of the values of individuals in the database. The weighting is according to similarity.

Attribute similarity solution to the reference class problem on basis of raw data. The direct inference probability that c has attribute A is given by

$$P(Ac) = T \sum_{i=1}^{n} sim_{Gow}(b_i, c) A(b_i) \quad , \tag{4.4}$$

where T is a normalising constant.

For instance, let Bc, Cc, Hc, Ec and

1. $Ab_1, \neg Eb_1, Cb_1$
2. $Ab_2, Bb_2, \neg Cb_2, Eb_2, Hb_2$
3. $\neg Ab_3, Bb_3, Cb_3$

then $Prob(Ac) = \frac{4}{7}(\frac{1}{2} \cdot 1 + \frac{3}{4} \cdot 1 + \frac{2}{2} \cdot 0) = \frac{5}{7}$.

This approach addresses both the above obstacles that face generic-probability approaches. In order to apply this method, one does not need a large sample of individuals who have a particular combination of attributes—we may assign a direct inference probability without considering frequencies in reference classes at all. Data in different datasets measuring different variables can be employed without worrying about the interaction of these variables; all the work is done by the similarity measure. Hence, no inconsistent marginals arise. Therefore, this single-case reinterpretation of the similarity approach apparently circumvents both the problem of small sample size and the problem of inconsistent marginal probabilities.

One might object to this apparent resolution as follows. Perhaps the single-case version of the similarity approach, although applicable, should not be applied when there is insufficient data, because, when there is insufficient data, it is not reasonable to infer anything about the individual. In response to this objection, note that the single-case version of the similarity approach is based on the assumption that the

best way to determine the direct inference probability for a certain individual is to study individuals that have similar attributes. According to this approach, to which reference class these individuals belong is irrelevant; what is relevant is that they are similar to the individual in question. Therefore, data that is insufficient for estimating generic probabilities may yet be sufficient for direct inference. Indeed, single-case versions of the similarity approach have been successfully used in machine learning and medicine—see Jerez et al. (2010) and Davis et al. (2010).

4.6 A Mechanism-Based Approach

In this section, we identify two loci where evidence of mechanisms may be used to provide better solutions to the reference class problem. First, evidence of mechanisms may be used to enrich the problem formulation, to better capture the causal structure of the direct inference problem in question. Capturing more features of the problem promises to lead to more accurate direct inferences. Second, evidence of mechanisms may be used to provide information about reference class probabilities for which we have no statistical information available but which are relevant to the direct inference at hand. The method of *comparative process tracing* employs evidence of similarity of mechanisms to extrapolate frequencies from a reference class for which we have data to one that is relevant to the direct inference.

In this section, we explain the concept of mechanism and discuss these two situations in which evidence of mechanisms can help solve reference class conflicts.

4.6.1 Evidence of Mechanisms and Causal Structure

Illari and Williamson characterise mechanism in the following way: "A mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon" (Illari and Williamson 2012, p. 120). Examples of mechanisms are the mechanism for drug metabolism in humans, the mechanism of natural selection and the mechanism of how supernovae arise. For instance, Russo and Williamson (2007, p. 162) describe the mechanism that leads from smoking to cancer by "The hair-like cilia in the lungs, which beat rhythmically to remove inhaled particles, are destroyed by smoke inhalation; thus the lung cannot cleanse itself effectively. Cancer-producing agents in cigarette smoke are therefore trapped in the mucus. Cancer then develops when these chemical agents alter the cells, in particular, cell division." According to Russo and Williamson, mechanisms play a crucial role in establishing causality. To establish that an event C causes an event E, normally two claims have to be established: that C and E are probabilistically dependent conditional on other causes of E, and that there exists a mechanism connecting C and E that can account for this correlation. One way of establishing that a certain mechanism connects C and E is to establish the crucial attributes of

the mechanism, i.e., to establish that the crucial entities and activities are present and are organized in the right way.

How we can get evidence of a mechanism and what counts as high quality evidence of mechanism is an active area of research (for an overview see Clarke et al. 2014). Evidence of mechanisms can come from various sources, including laboratory experiments, literature reviews of basic science, expert testimony, confirmed theory or by analogy from, e.g., animal experiments (Clarke et al. 2014).

In what is to follow, instead of speaking of a single mechanism, we are going to speak of the *mechanistic structure* that gives rise to an attribute measured in a reference class. The mechanistic structure consists of all the mechanisms that explain the attribute, either by inducing the attribute or by inhibiting it or by moderating its value (i.e., by changing its value or limiting change that would otherwise occur).

Evidence of mechanisms can assist direct inference by providing information about causal structure. As we have seen, in direct inference we seek to ascertain the probability that a particular individual instantiates a particular attribute. Statistical information is normally available, which takes the form of the generic probability of the attribute of interest conditional on other attributes which define a reference class. It is usually the case that there is also information to hand about the mechanisms that give rise to the attribute of interest and that connect this attribute to those that define the reference class. That this information about mechanisms is often qualitative rather than quantitative does not, of course, imply that it should be ignored for the purposes of direct inference. Evidence of mechanisms can be taken into account by helping to ascertain the causal structure that connects the attribute of interest to those attributes that characterise the reference classes for which we have statistics.

To the extent that this causal and statistical information underdetermines the required direct inference probability, one can apply objective Bayesian methods to select a direct inference probability that satisfies the constraints imposed by the available evidence but which is otherwise equivocal.

Constraints on the objective Bayesian probability function arising from the available statistical information can be dealt with by the approach introduced in Sect. 4.2. But now there are also causal constraints, which arise from evidence of mechanisms in combination with available statistical information. Williamson (2005a, §5.8) describes how causal constraints can be taken into account by the objective Bayesian approach. Briefly, objective Bayesianism adopts the principle that, when one learns of new variables which are not causes of the old variables, probabilities over the old variable set should not change. This principle allows one to translate information about causal relationships into constraints that equate probabilities. One can then identify the probability function $P^\dagger$ with maximum entropy, from all those probability functions which satisfy the constraints arising from statistical information together with the equality constraints arising from causal information. As described in Sect. 4.2, $P^\dagger$ is used for direct inference.

The mechanistic approach involves enriching the problem formulation by taking causal structure and extra causes and effects into account. The question arises, then, as to whether there are methods for mitigating this extra complexity. Maximis-

ing entropy is a computationally demanding optimisation problem, and practical methods are required to ensure that the optimisation can be carried out. Bayesian networks are often used to reduce the complexity of probabilistic inference, and they can also be applied here, as we shall now explain.

The probability function $P^\dagger$ advocated by objective Bayesianism can be represented by a Bayesian network—this is called an *objective Bayesian net* (Williamson 2005b; Landes and Williamson 2016). A Bayesian net consists of a directed acyclic graph whose nodes are variables together with the probability distribution of each variable conditional on its parents in the graph (Pearl 1988). The *Markov condition*, which says that each variable is probabilistically independent of its non-descendants in the graph conditional on its parents, then enables the Bayesian net to completely determine a joint probability distribution over all the variables that occur in the net. It turns out that, in virtue of its appeal to maximum entropy methods, the objective Bayesian probability function $P^\dagger$ typically satisfies many probabilistic independence relationships and these relationships can be identifiable in advance, to build an objective Bayesian net representation of $P^\dagger$ (Williamson 2005a, §§5.6–5.8). The advantages of the Bayesian net representation are that (i) it is a more economical way to specify a probability function than simply specifying the probability of each combination of values of the variables under consideration, and (ii) a whole host of efficient inference algorithms have been developed to calculate conditional probabilities—such as are required for direct inference—from the network.

Note that in general, the arrows in the graph of a Bayesian net are merely a formal device to represent certain probabilistic independencies—they would not normally be interpretable as representing causal relationships or other sorts of relationship. However, it turns out that in many situations where causal information is available, the objective Bayesian net is interpretable as a *causal* Bayesian net, i.e., the arrows in the directed acyclic graph represent causal relationships amongst the variables. This is so, for example, when the evidence consists of the causal structure and constraints on the probability distribution of each variable conditional on its parents; then the maximum entropy probability function $P^\dagger$ that is advocated by objective Bayesianism is determined by a causal net involving that causal structure—i.e., the Markov condition provably holds (Williamson 2005a, Theorem 5.8). So, in many situations the objective Bayesian approach can be thought of as a causal network approach.

In sum, the mechanistic approach to the reference class problem motivates two developments to the objective Bayesian approach. First, causal constraints need to be taken into account, in addition to probabilistic constraints. Second, Bayesian networks can be used to make direct inference more computationally tractable.

Let us consider an example. Bernd is a white man from the US who has never been to hospital. Bernd smokes and is highly physically active. We are interested in whether Bernd will get a stroke (St). Bernd belongs to two reference classes for which we have data: he smokes (S), and he is physically active (A). On the one hand, smoking increases the risk of getting a stroke by a half, $P^*(St|S) = 1.5 \cdot P^*(St|\neg S)$ (Shinton and Beevers 1989). On the other hand, a high degree of physical activity

decreases risk of stroke compared to a low degree of physical activity, $P^*(St|A) = 0.79 \cdot P^*(St|\neg A)$ (Mozaffarian et al. 2015). Belonging to the two reference classes yields opposite conclusions as to whether Bernd will get a stroke. What is the direct inference probability that Bernd will get a stroke? Mechanistic evidence is useful to constrain the direct inference probability and for an improved understanding of the causal structure of the problem. An improved understanding of the causal structure can yield tighter constraints on the direct inference probabilities, and can thus lead to direct inference probabilities that are better calibrated to the true chances.

The mechanism-based approach proceeds as follows. As a first step, we use available information about the mechanisms for stroke that involve smoking and physical activity and use this information to construct a causal net which schematically represents the causal relationships connecting these variables. In examples such as this, there is usually plenty of information about mechanisms available in the literature. Indeed, there are many mechanisms connecting smoking to stroke. Smoking increases the risk of building blood clots and blood clots increase the risk of stroke. Smoking increases the risk of high blood cholesterol (C) and high cholesterol increases the risk of some kinds of stroke. Similarly, there are many inhibiting mechanisms connecting physical activity to stroke. Physical activity reduces obesity and obesity is a major risk factor for stroke. Physical activity prevents hypertension, high cholesterol and the development of blood clots. Hypertension increases the risk of stroke.

The aim of the present example is to illustrate the mechanism-based approach, rather than to provide a detailed analysis of all mechanisms relating stroke, physical activity, and smoking. We therefore simplify the example by taking the mechanism involving high cholesterol to be the only mechanism that is influenced by both smoking and physical activity. Clearly, the example can be further enriched to take into account other evidence of mechanisms, and to lead to further improvements in direct inference.

Employing mechanistic evidence translates into the causal structure depicted in Fig. 4.1. Both smoking and physical activity influence cholesterol levels. Therefore, we draw an arrow from S to C and from A to C in the causal graph. Since high cholesterol causes stroke, we draw an arrow from C to St. Smoking and physical activity influence stroke via at least two different non-overlapping mechanisms. Therefore, we draw arrows from both S and A to St.

As a second step, we need to ascertain any other available probabilities that are relevant to the variables that occur in the causal graph. These probabilities can often

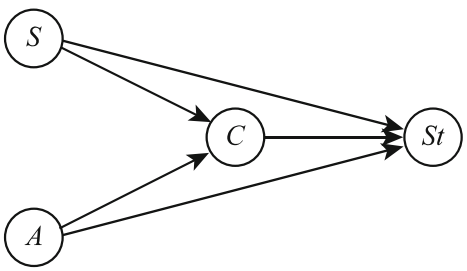

Fig. 4.1 Causal graph for the stroke example

also be found in the literature. In order to determine $P^*(St|S \wedge A)$, it would suffice to specify, for every variable appearing in the graph, the probability distribution of that variable conditional on its parents in the graph. This would constitute a causal Bayesian net, which would fully determine the joint chance distribution over the variables that occur in the graph.

However, it will typically be the case that not all of these probabilities are available in the literature. Instead, the conditional probability of a variable given some of its parents might be available. In this case, we treat these "incomplete distributions" as constraints and carry out maximum entropy direct inference, as described above.

In the literature, relative risks are often reported. For instance, $P^*(St|S) = 1.5 \cdot P^*(St|\neg S)$ (Shinton and Beevers 1989). In this case, we obtain the absolute risk of stroke given smoking $P^*(St|S)$, if we know the base rate or prevalence of stroke and hypertension, $P^*(St)$ and $P^*(S)$:

$$\begin{aligned} P^*(St) &= P^*(St|S)P^*(S) + P^*(St|\neg S)(1 - P^*(S)) \\ &= 1.5 \cdot P^*(St|\neg S)P^*(S) + P^*(St|\neg S)(1 - P^*(S)) \end{aligned}$$

From the literature, we obtain the prevalence or base rate of stroke $P^*(St) = 0.027$, of smoking $P^*(S) = 0.168$, of high cholesterol $P^*(C) = 0.131$ and of physical activity $P^*(A) = 0.695$ (Mozaffarian et al. 2015). Hence, $P^*(St|S) = 0.032$ and $P^*(St|A) = 0.025$.

Estimates of the risk of stroke given high cholesterol compared to the case of no high cholesterol are more controversial:

> The role of blood cholesterol in stroke prevention is unclear. Most prospective studies have failed to find a relation between total cholesterol and risk of total stroke. It has been proposed that this may be due to the differing association with subtypes of stroke. An inverse association has been observed with hemorrhagic strokes and a positive association with ischemic stroke. (Wannamethee et al. 2000, p. 1887)

We differentiate between ischemic stroke ($St = 1$) and haemorrhagic stroke ($St = 2$). We abbreviate $St = 1 \vee St = 2$ by St. The following estimates can be found in the literature: $P^*(St = 1|C) = 1.4 \cdot P^*(St = 1|\neg C)$ (Benfante et al. 1994) and $P^*(St = 2|C) = 0.69 \cdot P^*(St = 2|\neg C)$ (Wang et al. 2013). The direct inference probability can be further constrained by, for instance, ascertaining $P^*(C|S)$, $P^*(C|A)$ and the high cholesterol-smoking interaction with respect to the development of stroke $P^*(St|C \wedge S)$.

As a third step, we can apply objective Bayesian methods to infer a direct inference probability. This probability is determined by the probability function with maximum entropy, from all those that satisfy the constraints imposed by the causal and statistical information, as explained above. This way, we obtain the direct inference probability that is compatible with the available evidence, but otherwise equivocal.

4.6.2 Evidence of Mechanisms and Extrapolation

Evidence of mechanisms can also be used to help extrapolate available reference class probabilities to classes for which we have no statistical information available but which are relevant to the direct inference at hand. Thus, a direct inference probability can be influenced by frequencies obtained from reference classes to which the individual does not belong.

Broadly speaking, in order to extrapolate a probabilistic claim from one reference class to another, one needs to show that the determinants of the probability in the target class are sufficiently similar to those in the source class. One can do this by showing that the underlying mechanisms, contextual factors and background conditions are sufficiently similar in the two classes. Clearly, evidence of mechanisms is crucial to this mode of inference.

There are various strategies of employing evidence of mechanisms to assist extrapolation (Parkkinen and Williamson 2017). *Comparative process tracing* is one such strategy (Steel 2008). To determine how likely an extrapolation from a model organism to a target organism is to succeed, Steel proposes that one has to learn the mechanism in the model organism and that one has to compare stages of the mechanism in which the mechanism in the model and the target are most likely to differ significantly. He then concludes that "in general, the greater the similarity of configuration and behavior of entities involved in the mechanism at these key stages, the stronger the basis for the extrapolation" (Steel 2008, p. 89). Steel illustrates his method by means of an example. According to Steel, rats are better models than mice for determining whether Aflatoxin B1 causes liver cancer in humans. He argues that (i) phase 1 and phase 2 metabolism are the crucial parts in most carcinogenic mechanisms among mammals and that (ii) the phase 1 metabolism is similar in all three species and that (iii) there are important similarities between rats and humans in phase 2 metabolism but dissimilarities between humans and mice. Therefore, according to comparative process tracing, the extrapolation from rats to humans is more likely to succeed.

Consider now the prevalence of smoking in a country, state or city. Can we extrapolate the chance to another country, state or city? This depends on how similar the crucial attributes or determinants of the smoking behaviour are in the target and the study population. For instance, demographic or socioeconomic factors are important determinants of smoking behaviour: e.g., education, income level of the country, age, and gender (Hosseinpoor et al. 2011). It is legitimate to extrapolate the prevalence of smoking from one state in a high-income country to another state, provided that there is little difference in socioeconomic or demographic factors. For instance, in Austria, it is reasonable to extrapolate the chance of smoking from the State of Tirol to the State of Lower Austria. Indeed, the smoking rates in Austria in 7 out of 9 states differ by less than 2% (20.9%–22.7%) (Statistik Austria 2016). It is, however, not reasonable to extrapolate from the State of Tirol to the State of Vienna. While the State of Vienna consists roughly of the large city, the State of Tirol consists mainly of rural areas and smaller cities.

Often it is impossible to avoid extrapolation in direct inference, because statistics in the relevant reference classes are not available. Thus in the stroke example, in order to specify probabilities relevant to the causal net, we extrapolated estimates for the probabilities in the study population to the population to which Bernd belongs. For instance, since Bernd has never been hospitalised, Bernd does not belong to any of the hypercholesterol populations from which the samples were drawn—these studies were conducted on hospital patients. Less straightforward extrapolation has been carried out to obtain the risk of ischemic stroke given high blood cholesterol. The estimate $P^*(St = 1|C) = 1.4 \cdot P^*(St = 1|\neg C)$ from Benfante et al. (1994) is obtained from studies conducted on Hawaiian Japanese men. There are surely important differences between Hawaiian Japanese men and US-born citizens. However, "the associations of major risk factors with CHD and stroke were very similar to those found for US white men" (Benfante et al. 1994, p. 818). This provides grounds that an extrapolation from the risk of ischemic stroke to Bernd's population (US white men) will be successful, i.e., that $P^*(St = 1|C) \approx 1.4 \cdot P^*(St = 1|\neg C)$ in US white men.

4.7 Conclusions

We have presented four approaches to the reference class problem. The objective Bayesian approach advocates degrees of belief that are compatible with the evidence and otherwise equivocal. Pollock's approach is based on combinatorial probabilities. The similarity approach is based on similarity of attributes or of individuals. Finally, the mechanistic approach allows one to enrich and refine the problem formulation, in order to achieve more credible direct inferences. Which of the approaches should we prefer?

To give an answer, at least three further questions have to be considered. First, how is the aggregation of reference classes done? The similarity approach identifies the direct inference probability with a weighted average of reference class frequencies. This weighted average seems to be somewhat arbitrary: why should we combine different evidence exactly in this way? The objective Bayesian approach and Pollock's approach relate the direct inference to less ad hoc quantities: respectively, the maximum entropy probability for the narrowest reference class to which the individual is known to belong, and the expectable value of the narrowest reference class. We argued above, however, that Pollock's justification of the claim that the expectable value is often close to the true chance is unconvincing. Advocating equivocal degrees of belief on the basis of good but incomplete evidence is preferable to advocating degrees of belief on the basis of complete but poor evidence. Hence, aggregation according to the objective Bayesian approach might be considered preferable to aggregation via the other two approaches.

Second, what kind of evidence can be used? The attribute-based similarity measure does not account for causal similarity. Generally, measures based on the number of shared attributes can only be seen as a first approximation to a similarity

measure that will be useful for direct inference. The objective Bayesian account can account for causal information by exploiting evidence of mechanisms. This leads to tighter empirical constraints on the direct inference probability. Of course, that it is able to exploit evidence of mechanisms is only an advantage if there is evidence of mechanisms available that is of sufficiently high quality.

Third, the similarity approach can be reconstructed as a single-case-probability approach and may be used even when the sample size is too small to reliably estimate the data-generating chances. The strategy of taking into account similarity of individuals might thus compensate for low sample size.

Our suggestion is to prefer the similarity approach where sample sizes are too small to sensibly apply generic-probability approaches. In the other cases, the objective Bayesian approach is to be preferred. This is especially true if evidence of mechanisms is available that constrains the direct inference probability to lie in a tight interval. In such a case, there is less of a role for the Equivocation norm and most of the work is being done by the Calibration norm. This is likely to lead to accurate, rather than cautious, direct inference probabilities.

Acknowledgements The research for this paper has been funded by the Arts and Humanities Research Council via grant AH/M005917/1.

References

Benfante, R., K. Yano, L.-J. Hwang, J.D. Curb, A. Kagan, and W. Ross. 1994. Elevated serum cholesterol is a risk factor for both coronary heart disease and thromboembolic stroke in hawaiian Japanese men. Implications of shared risk. *Stroke* 25(4): 814–820.

Clarke, B., D. Gillies, P. Illari, F. Russo, and J. Williamson. 2014. Mechanisms and the evidence hierarchy. *Topoi* 33(2): 339–360.

Davis, D.A., N.V. Chawla, N.A. Christakis, and A.-L. Barabási. 2010. Time to care: A collaborative engine for practical disease prediction. *Data Mining and Knowledge Discovery* 20(3): 388–415.

Gower, J.C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27(4): 857–871.

Haybittle, J., R. Blamey, C. Elston, J. Johnson, P. Doyle, F. Campbell, R. Nicholson, and K. Griffiths. 1982. A prognostic index in primary breast cancer. *British Journal of Cancer* 45(3): 361.

Hosseinpoor, A.R., L.A. Parker, E.T. d'Espaignet, and S. Chatterji. 2011. Social determinants of smoking in low-and middle-income countries: Results from the world health survey. *PLoS One* 6(5): e20331.

Illari, P.M., and J. Williamson. 2012. What is a mechanism? Thinking about mechanisms *across* the sciences. *European Journal for Philosophy of Science* 2: 119–135.

Jaynes, E.T. 1957. Information theory and statistical mechanics. *The Physical Review* 106(4): 620–630.

Jerez, J.M., I. Molina, P.J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco. 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine* 50(2): 105–115.

Landes, J., and J. Williamson. (2016). Objective Bayesian nets from consistent datasets. In *Proceedings of the 35th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. A. Giffin, and K.H. Knuth, vol. 1757. American Institute of Physics Conference Proceedings, Potsdam

Mozaffarian, D., E.J. Benjamin, A.S. Go, D.K. Arnett, M.J. Blaha, M. Cushman, S. de Ferranti, J.-P. Despres, H.J. Fullerton, V.J. Howard, et al. 2015. Heart disease and stroke statistics-2015 update. A report from the American Heart Association. *Circulation* 131(4): E29–E322.

Parkkinen, V.-P., and J. Williamson. 2017. Extrapolating from model organisms in pharmacology. In *Uncertainty in pharmacology: Epistemology, methods, and decisions*, ed. B. Osimani. Dordrecht: Springer.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo: Morgan Kaufmann.

Pollock, J.L. 2011. Reasoning defeasibly about probabilities. *Synthese* 181(2): 317–352.

Pollock, J.L. 2016. LISP code for OSCAR. http://johnpollock.us/ftp/OSCAR-web-page/oscar.html. Accessed 19 Sept 2016.

Russo, F., and J. Williamson 2007. Interpreting causality in the health sciences. *International Studies in the Philosophy of Science* 21(2): 157–170.

Shinton, R., and G. Beevers. 1989. Meta-analysis of relation between cigarette smoking and stroke. *BMJ* 298(6676): 789–794.

Statistik, Austria. 2016. Smoking statistics in Austria 2014. http://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/gesundheit/gesundheitsdeterminanten/rauchen/index.html. Accessed 28 Sept 2016.

Steel, D. 2008. *Across the boundaries: Extrapolation in biology and social science*. Oxford/New York: Oxford University Press.

Venn, J. 1888. *The logic of chance*, 3rd ed. London: Macmillan.

Wallmann, C. 2017. A Bayesian solution to the conflict of narrowness and precision in direct inference. *Journal for General Philosophy of Science*.

Wallmann, C., and G.D. Kleiter 2014a. Degradation in probability logic: When more information leads to less precise conclusions. *Kybernetika* 50(2): 268–283.

Wallmann, C., and G.D. Kleiter 2014b. Probability propagation in generalized inference forms. *Studia Logica* 102(4): 913–929.

Wang, X., Y. Dong, X. Qi, C. Huang, and L. Hou. 2013. Cholesterol levels and risk of hemorrhagic stroke. A systematic review and meta-analysis. *Stroke* 44(7): 1833–1839.

Wannamethee, S.G., A.G. Shaper, and S. Ebrahim. 2000. HDL-cholesterol, total cholesterol, and the risk of stroke in middle-aged British men. *Stroke* 31(8): 1882–1888.

Williamson, J. 2005a. *Bayesian nets and causality: Philosophical and computational foundations*. Oxford: Oxford University Press.

Williamson, J. 2005b. Objective Bayesian nets. In *We will show them! essays in Honour of Dov Gabbay*, ed. S. Artemov, H. Barringer, A.S. d'Avila Garcez, L.C. Lamb, and J. Woods, vol. 2, 713–730. London: College Publications.

Williamson, J. 2010. *In defence of objective Bayesianism*. Oxford: Oxford University Press.

Williamson, J. 2013. Why frequentists and Bayesians need each other. *Erkenntnis* 78(2): 293–318.

Williamson, J. 2017a. *Lectures on inductive logic*. Oxford: Oxford University Press.

Williamson, J. 2017b. Models in systems medicine. *Disputatio*. In press.