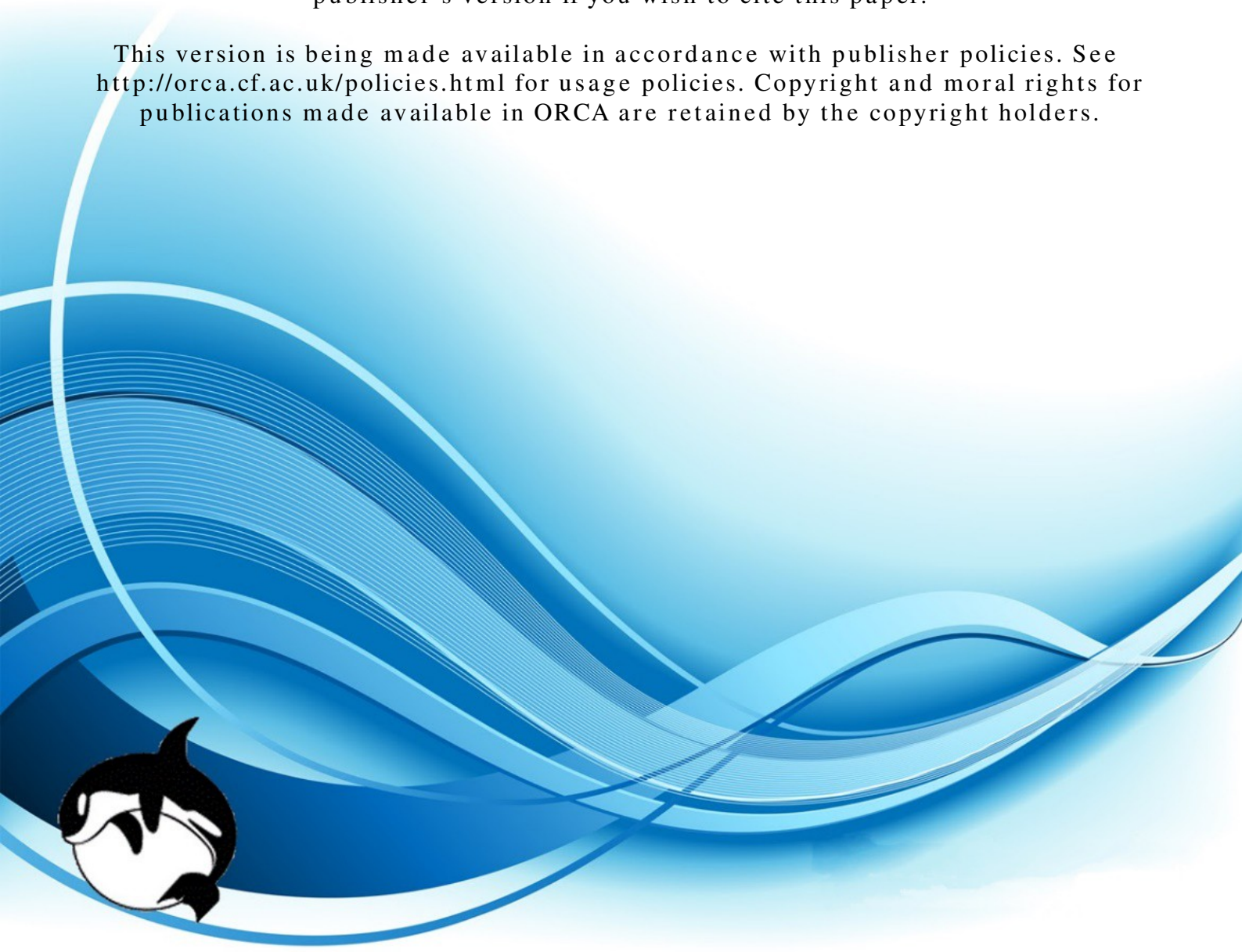

Statistical Methods in Medical Research
XX(X):1–9

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Understanding between-cluster variation in prevalence, and limits for how much variation is plausible

Mark D. Chatfield[1] and Daniel M. Farewell[2]

Abstract

In clinical trials and observational studies of clustered binary data, understanding between-cluster variation is essential: in sample size and power calculations of cluster randomised trials, for example, the intra-cluster correlation coefficient (ICC) is often specified. However, quantifications of between-cluster variation can be unintuitive, and an ICC as low as 0.04 may correspond to surprisingly large between-cluster differences. We suggest that understanding is improved through visualising the implied distribution of true cluster prevalences—possibly by assuming they follow a beta distribution—or by calculating their standard deviation, which is more readily interpretable than the ICC. Even so, the bounded nature of binary data complicates the interpretation of variances as primary measures of uncertainty, and entropy offers an attractive alternative. Appealing to maximum entropy theory, we propose the following rule of thumb: that plausible ICCs and standard deviations of true cluster prevalences are both bounded above by the overall prevalence, its complement, and one third. We also provide corresponding bounds for the coefficient of variation (CV), and for a different standard deviation and ICC defined on the log odds scale. Using previously published data, we observe the quantities defined on the log odds scale to be more transportable between studies with different outcomes with different prevalences than the ICC and CV. The latter increase and decrease, respectively, as prevalence increases from 0% to 50%, and the same is true for our bounds. Our work will help clinical trialists better understand between-cluster variation, and avoid specifying implausibly high values for the ICC in sample size and power calculations.

Keywords

Intra-cluster correlation coefficient, Intra-class correlation coefficient, coefficient of variation, prevalence, proportion, binary outcome, cluster randomised trial, sample size, maximum entropy, bounds

Introduction

Quantifying between-cluster or between-site variation is of interest in observational studies[1] and in randomised controlled trials[2,3], especially cluster randomised trials. In a parallel-group cluster randomised trial, clusters such as healthcare organisations, school classes or geographic areas are randomised to trial arms (e.g. intervention or control), and outcomes are measured on individuals within those clusters. Less common types of cluster randomised trials include stepped-wedge designs and cluster-crossover designs.

In sample size and power calculations for cluster randomised trials, the between-cluster variation in the primary outcome needs to be anticipated, and in a parallel-group cluster randomised trial, this variation must be anticipated for each arm. For trials with a binary primary outcome, this is nearly always done by specifying a single value of the intra-cluster (or intra-class) correlation coefficient (ICC)[2,4]. Cluster randomised trials require more individuals than individually randomised controlled trials: a parallel-group cluster randomised trial should be $1 + (m - 1) \times \text{ICC}$ times larger than the corresponding individually randomised controlled trial, where m is the average number of individuals in a cluster[2].

The ICC quantifies homogeneity of outcome within clusters, and can be expressed as the proportion of the total variance that is accounted for by between-cluster rather than within-cluster variation. Perhaps more helpfully, the ICC is also the correlation between the outcomes of any pair of individuals in the same cluster. For binary outcome data, the ICC is equivalent to the kappa statistic[5].

None of the aforementioned ways of expressing the ICC lend themselves to helping researchers guage whether a certain value of the ICC is "high" or "low". As the ICC ranges from 0 (no variation in cluster prevalences) to 1 (all individuals within a cluster have identical outcomes), a researcher might understandably guess that an ICC of 0.04 (a fairly typical value) corresponds to a tiny amount of between-cluster variation.

Making things harder to understand still, empirically observed ICCs vary with the prevalence of the binary outcome[5], and there is an alternative definition of the ICC, defined not on the probability scale but on the log odds scale, which can differ considerably[6]. It is not hard for researchers new to cluster randomised trials involving binary outcomes to confuse the two definitions. Other definitions of the ICC

[1]The University of Queensland, Australia
[2]Cardiff University, UK

Corresponding author:
Daniel Farewell, Division of Population Medicine, School of Medicine, Cardiff University, Cardiff, CF14 4YS, UK
Email: farewelld@cardiff.ac.uk

also exist, for example derived from models that adjust for covariates[6].

The ICC has been described as unintuitive and harder to understand than the coefficient of variation (CV) of the cluster-specific prevalences, which is an alternative way of describing the between-cluster variation[7]. Sample size for a cluster randomised trial is sometimes calculated for a binary outcome using the CV[8]. The calculation is not as neat as the one using the ICC, and care must be taken if the anticipated prevalence exceeds 50%.

Another motivation for the CV is that it may be 'transportable' between groups: if a treatment acts multiplicatively on the risk of an adverse event, then the CV will be the same in both intervention and control groups[7]. Crespi et al.[9] offer similar motivation (transportable "across study conditions and between studies with different outcome prevalences") for another measure of between-cluster variation, $R = 1 + \text{ICC} \times (1-\mu)/\mu$, where μ is the overall prevalence. However, R can equivalently be written as $1 + \text{CV}^2$[10], and hence is transportable if and only if the CV is.

In observational studies, there are many ways of quantifying between-cluster variation[1,11]. Most express the difference between two clusters using an odds ratio: a cluster mean that exceeds another on the log odds scale by σ_L will have larger odds by a factor of $\exp(\sigma_L)$. This may make it easier to compare (possibly residual) between-cluster variation with the effects of covariates (also expressed as odds ratios) and ensures that the between-cluster variation does not depend on covariates, but it requires some mental gymnastics to determine what this variation actually corresponds to on the probability scale. The easiest way is probably to calculate a few percentiles, for example the 2.5th and 97.5th percentiles, for one or more fixed prevalences.

In this paper we aim to help researchers in their understanding of between-cluster variation in studies with binary outcomes. We also describe what we determine to be the largest plausible values of several common measures of between-cluster variation.

Notation

We shall think of cluster-specific true prevalences as random effects[6]. We denote by p_i the true (rather than observed) prevalence of a binary outcome in cluster i, and let $\mu = E(p_i)$ and $\sigma = \text{SD}(p_i)$ be their mean and standard deviation, respectively. While we typically think of p_i representing the probability of an adverse event, there is no reason not to work instead in terms of $q_i = 1 - p_i$, the probability of *avoiding* an adverse event. (The only measure of variability for which this distinction makes a difference is the coefficient of variation.) We sometimes use the phrase "overall prevalence" in place of "mean cluster prevalence", implicitly assuming that all clusters are in principle infinite. The coefficient of variation of the cluster prevalences is

$$\text{CV}(p_i) = \frac{\text{SD}(p_i)}{\text{E}(p_i)} = \frac{\sigma}{\mu}.$$

If y_{ij} is a binary outcome on individual j in cluster i, Eldridge et al.[6] define the ICC as

$$\text{ICC} = \frac{\text{Var}(p_i)}{\text{Var}(y_{ij})} = \frac{\sigma^2}{\mu(1-\mu)}.$$

and an alternative ICC (on the log odds scale) as

$$\text{ICC}_L = \frac{\sigma_L^2}{\sigma_L^2 + \pi^2/3},$$

where

$$\sigma_L = \text{SD}(\text{logit}\, p_i)$$

is the standard deviation of the cluster prevalences on the unconstrained log odds scale, with $\text{logit}\, p_i = \log\{p_i/(1-p_i)\}$ denoting the usual logistic transformation.

This alternative definition of ICC is motivated by the fact that a binary outcome may be thought of as a dichotomised version of an underlying continuous variable: individuals for whom the value of this latent continuous variable is above a certain threshold have a value of 1 for the binary outcome, and all other individuals have a value of 0. If, given cluster prevalences, the underlying latent continuous variable is assumed to follow a standard logistic distribution (whose variance is $\pi^2/3$), then the ICC on the log odds scale can be interpreted as the proportion of the total variance of the latent continuous variable that is between, as opposed to within, clusters. In this way, ICC_L is made analogous to an ICC used for continuous outcomes.

Although this alternative, log odds definition of the ICC implies a residual logistic distribution given cluster means, none of the foregoing measures of between-cluster variation depend on assuming any particular parametric form for the distribution of the cluster prevalances p_i. Nevertheless, such parametric assumptions can be helpful for visualisation or computation, and we now discuss several possible choices.

Logistic-normal distribution

With an unspecified distribution on the cluster prevalences, the alternative ICC (ICC_L) is implicitly invoking a generic logistic mixture model. The logistic-normal model is a generalised linear mixed model[12] that incorporates a Gaussian mixing distribution and is commonly applied in trials and in observational studies. While cluster-specific factors such as trial arm are allowed to affect μ, the logistic-normal model assumes that the $\text{logit}\, p_i$ have a normal distribution and that their standard deviation σ_L does not change with cluster-specific factors or other covariates.

Figure 1 shows the between-cluster variation implied by a logistic-normal model with $\sigma_L = 0.4$ (and hence $\text{ICC}_L = 0.046$) for 5 different prevalences ranging from $\mu = 0.1$ to $\mu = 0.9$. Despite the fact that σ_L is held constant, σ gets smaller as μ gets further from 0.5, and at a particularly fast rate as μ approaches 0 or 1. The same is true of the ICC, which takes a value slightly smaller than ICC_L at $\mu = 0.5$ ($\text{ICC} = 0.037 < 0.046$), but which is considerably smaller than ICC_L at $\mu = 0.1$ ($\text{ICC} = 0.015 \ll 0.046$). As μ gets closer to 0, the CV increases.

Beta distribution

Like the normal distribution, the beta distribution is a continuous distribution determined by two parameters, often

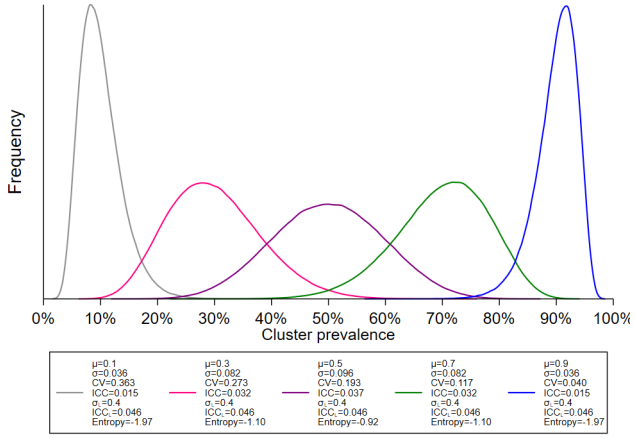

Figure 1. Distribution of cluster prevalences p_i when σ_L held constant while mean cluster prevalence μ varies. (Constructed following simulation of logistic-normal data, and smoothed using kernel density estimation.)

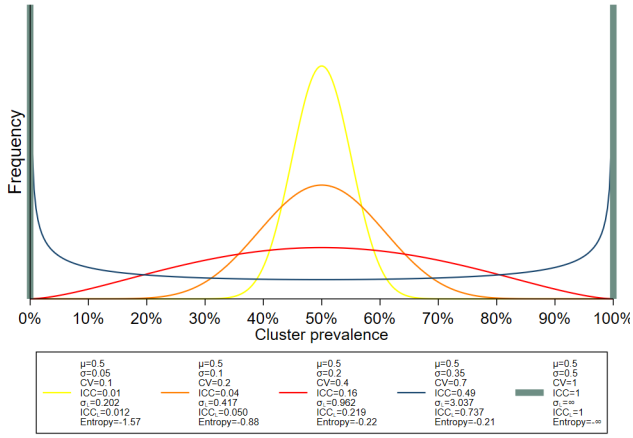

Figure 2. Distribution of cluster prevalances p_i when mean cluster prevalence $\mu = 0.5$. (Beta distribution assumed.)

labelled (α, β). Unlike the normal distribution, values are restricted to lie within the range 0 to 1, so the beta distribution is a convenient and flexible option for describing the distribution of cluster prevalences p_i.

Assuming a beta distribution, the mean cluster prevalence μ is

$$\frac{\alpha}{\alpha+\beta}$$

and the ICC[13] is

$$\frac{1}{1+\alpha+\beta}.$$

Conversely, the parameters (α, β) needed to obtain a specified mean and ICC (μ, ICC) are

$$\alpha = \mu \times \frac{1-\text{ICC}}{\text{ICC}}$$

$$\beta = (1-\mu) \times \frac{1-\text{ICC}}{\text{ICC}}.$$

With the right statistical software package, visualising a beta distribution is straightforward, as is calculating percentiles. Here, we have used Stata[14].

Going between the log odds and probability scales

Simulation of data is one way of translating between the log odds and probability scales, and we recommend simulation as a good tool in general for checking that the variation being described is in fact realistic in the relevant application area. However, it is sometimes convenient to relate these two scales more directly.

Murray and Murray[15] give an approximation linking σ and σ_L, namely

$$\sigma \approx \sigma_L \times \mu(1-\mu)$$

that assumes a logistic-normal model and is derived from a Taylor series expansion valid for small σ_L. However, seemingly unknown in the field of ICC research is an elegant, exact calculation that links σ_L to the usual standard deviation σ and mean μ via the trigamma function[16] by assuming instead that cluster prevalences follow a beta distribution with parameters (α, β):

$$\sigma_L^2 = \text{trigamma}(\alpha) + \text{trigamma}(\beta)$$

This identity follows directly from the fact that the logistic transformation of a beta distribution is equivalent to the logarithm of the ratio of two independent gamma distributions; $\text{trigamma}(\alpha)$ and $\text{trigamma}(\beta)$ are their respective logarithmic variances. The identity provides a mathematically precise connection between a beta distribution on the probability scale and the standard deviation σ_L on the log odds scale, and we use it to compute σ_L in Figure 2 and Figure 3.

Visualisation

Figure 2 shows five different distributions of cluster prevalences p_i. They all have the same mean, $\mu = 0.5$, but they have differing amounts of variation about the mean. For example, the orange distribution has standard deviation $\sigma = 0.1$. As the beta distribution has been assumed, we can calculate that the middle 95% of clusters have prevalences ranging from 31% to 69%, in this case encompassing approximately 2 standard deviations either side of the mean. The coefficient of variation is CV $= 0.2$, and the intra-cluster correlation is ICC $= 0.04$.

It may surprise some readers, as it surprised us, that such a seemingly low value of ICC (0.04) can be associated with such a large amount of between-cluster variation. In some cases such instincts may derive from inter-rater reliability studies, where ICCs up to 0.4 can be classified as poor[17].

The green distribution represents the extreme scenario where 50% of clusters have a prevalence of 0% and 50% of clusters have a prevalence of 100%. This leads to the maximum possible between-cluster standard deviation ($\sigma = 0.5$, which is only possible when $\mu = 0.5$). The ICC is 1; unlike the maximal $\sigma = 0.5$, this maximal ICC is possible for any $0 < \mu < 1$.

More generally, the maximum possible between-cluster variance is $\sigma^2 = \mu(1-\mu)$, occurring when $100(1-\mu)\%$ of clusters have a prevalence of 0% and $100\mu\%$ of clusters have a prevalence of 100%[18]. Since $\mu(1-\mu)$ is the denominator

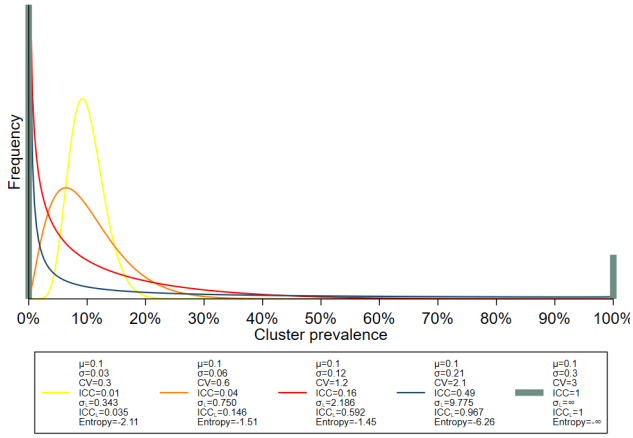

Figure 3. Distribution of cluster prevalances p_i when mean cluster prevalence $\mu = 0.1$. (Beta distribution assumed.)

in our definition of ICC, it follows that ICC can be expressed purely in terms of between-cluster variation: the ICC is the attained percentage of the maximum possible between-cluster variance, given mean cluster prevalence μ. To our knowledge, this is the first time the ICC has been expressed in this way. We argue that the extreme, discrete distribution with variance equal to this denominator variance is unrealistic in most applications, and may be one possible source of confusion about plausible magnitudes of ICCs.

Figure 3 shows five more distributions of cluster prevalances p_i. The ICCs are the same as those in Figure 2. The distributions in Figure 3 all have the same mean, $\mu = 0.1$, but they have differing amounts of variation about the mean. The (beta) distributions are not symmetric: the orange distribution has standard deviation $\sigma = 0.06$, and the middle 95% of clusters have prevalences ranging from 2% to 25%. The coefficient of variation is CV $= 0.6$, and the intra-cluster correlation is ICC $= 0.04$ as before. Once again, some readers may be surprised that a seemingly low value of ICC can be associated with such a large amount of between-cluster variation.

The values of ICC_L are slightly larger than corresponding values of ICC when $\mu = 0.5$ (Figure 2), but considerably larger when $\mu = 0.1$ (Figure 3).

Maximum entropy distribution

One reason our instincts about ICCs can mislead us is because of the temptation to confuse variance with uncertainty. Entropy is a more general notion of uncertainty that is applicable to all types of random variables, and "agrees with our intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one"[19]. For example, the degenerate green distributions of Figure 2 and Figure 3 have large variance but very small entropy, since only two values of p_i are possible. Put another way, though these distributions have maximal variance given their mean μ, there are many distributions with the same mean but larger entropy. A somewhat more detailed description of entropy may be found in an appendix to this paper.

It turns out that there is a uniquely determined density function that conveys the least information possible while remaining compatible with a particular prevalence μ. This is called the maximum entropy distribution and, as its name suggests, it has the largest entropy within this class of distributions[19]. Informally, it is the distribution that is the most "spread out" across its support, subject to the mean restriction that we choose to impose. It may be viewed as an extension of Laplace's principle of insufficient reason to cases where considerations of symmetry or uniformity are replaced by more general restrictions. From the perspective of information theory, the maximum entropy distribution is the least informative distribution within that class. The maximum entropy distribution therefore constitutes a natural reference distribution against which other distributions—all of which encode more information about the p_i—may be compared. Figure 4 shows four such maximum entropy distributions for various means, and again the appendix provides more detail about the mathematical derivation of maximum entropy distributions.

For a given prevalence μ, the standard deviation of the maximum entropy distribution is considerably smaller than the maximum possible standard deviation (Table 1). For example, for prevalences of 1%, 10% and 50%, the maximum possible standard deviations are, respectively, 0.1, 0.3 and 0.5, while the corresponding maximum entropy distribution standard deviations range from 0.01 to 0.29 ($= 1/\sqrt{12}$, the standard deviation of a uniform distribution on the unit interval). In agreement with our intuition, distributions with smaller standard deviation than that of the maximum entropy distribution also have lower entropy. Interestingly, increasing the standard deviation beyond that of the maximum entropy distribution has the effect of supplying additional information about the p_i and actually decreases uncertainty. This is a feature of the bounded range of the p_i, and a major reason why our intuition from studying ICCs for continuous responses (where entropy increases as between-cluster variation increases) can mislead us.

While considerations of plausibility are necessarily tentative, subjective and application-specific, we suggest that the distribution with largest entropy (for a given mean) helpfully bounds the range of standard deviations and ICCs that will arise in typical practice. For $\mu = 0.5$, the maximum entropy distribution is the uniform distribution on $(0, 1)$; for other values of μ, the maximum entropy distribution should be thought of as analogous to the uniform distribution, in the sense that the probability mass is dispersed as evenly as possible while respecting the mean restriction. As illustrated in Figure 4, for a general $0 \leq \mu \leq 1$ the maximum entropy distribution is a truncated exponential distribution (see Conrad[20], Theorem 5.1), with density proportional to $\exp(\lambda p_i)$ for a parameter λ. Unlike the distribution with largest variance for a given mean μ—which degenerates to provide support only on the two-point set $\{0, 1\}$—the corresponding maximum entropy distribution has a density that is strictly positive on the whole unit interval for every $0 < \mu < 1$. We suggest that the amount of between-cluster variation in real world examples will rarely be greater than the variation exhibited by this maximum entropy distribution.

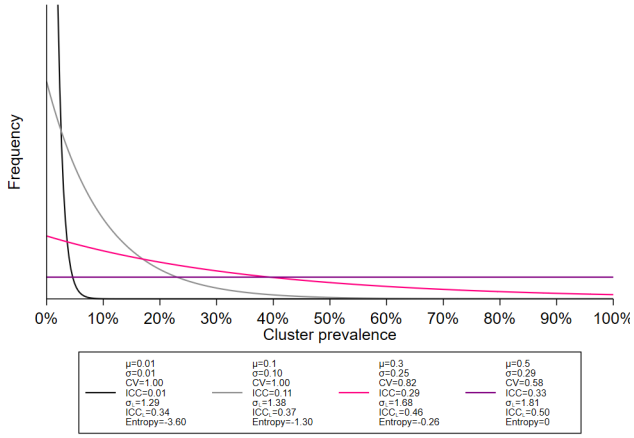

Figure 4. Distribution of cluster prevalances p_i according to four maximum entropy distributions (with mean cluster prevalences of $\mu = 0.01, 0.1, 0.3,$ and 0.5 respectively).

Table 1. Maximum plausible values and maximum possible values of measures of between-cluster variation in prevalence. The overall prevalence is μ. Plausible bounds are derived from the maximum entropy distribution.

	Plausible			Possible
Measure	$\mu \approx 0$	$\mu = 0.5$	$\mu \approx 1$	$0 \leq \mu \leq 1$
Probability scale:				
SD, σ	μ	0.29	$1-\mu$	$\sqrt{\mu(1-\mu)}$
CV	1	0.58	$(1-\mu)/\mu$	$\sqrt{(1-\mu)/\mu}$
ICC	$\mu/(1-\mu)$	0.33	$(1-\mu)/\mu$	1
Log odds scale:				
SD, σ_L	1.28	1.81	1.28	∞
ICC$_L$	0.33	0.50	0.33	1

Restricting attention to distributions with variance no larger than that of the maximum entropy distribution yields bounds on plausible ICCs and related quantities. These bounds vary with the mean: as σ in the maximum entropy distribution is smaller than μ for $\mu < 0.5$ and $\sigma < 1-\mu$ for $\mu > 0.5$ (Figure 5), it follows that ICC $< \mu/(1-\mu) \approx \mu$ for $\mu \ll 0.5$ and ICC $< (1-\mu)/\mu \approx 1-\mu$ for $\mu \gg 0.5$ (Table 1). Further, the ICC of the maximum entropy distribution when $\mu = 0.5$ (the uniform distribution) has ICC $= 1/3$. Hence a rough rule of thumb is this: plausible ICCs and cluster prevalence standard deviations are both bounded above by the overall prevalence, its complement, and one third. Figure 5 shows that our rule of thumb is satisfactory, even if it slightly underestimates the ICC arising from the maximum entropy distribution for $\mu < 0.25$ and $\mu > 0.75$.

Eldridge and Kerry[2] considered bounds for the ICC based on the beta distribution:

> "The most likely shape of this distribution is unimodal (having one peak) rather than U-shaped with peaks at 0 and 1, or J-shaped with a peak at either 1 or 0. Assuming a unimodal beta distribution, it is possible through algebraic manipulation of formulae representing the shape of the distribution to calculate the maximum possible ICC for different overall prevalences ... ICCs over 0.35 are unlikely for binary outcomes, and for extreme prevalences ICCs may be even smaller. Nevertheless, high values of the ICC may be observed in some trials as a result of sampling error, and in rare cases when the distribution of ICCs may not be unimodal."

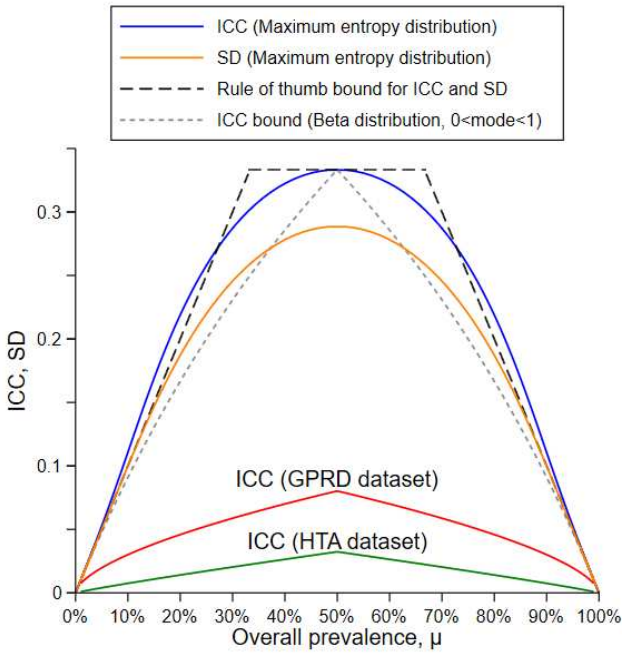

Figure 5. Bounds for the ICC and standard deviation describing the between-cluster variation in prevalence on the probability scale (assuming the maximum entropy distribution describes the maximum plausible amount of variation).

Their bound for the ICC based on the unimodal beta distribution (where the mode lies strictly between 0 and 1, i.e. $\alpha, \beta > 1$) is lower than our bound for the ICC based on the maximum entropy distribution except when $\mu = 0, 0.5$ or 1, when it is the same (Figure 5).

The CV associated with the maximum entropy distribution is near 1 when the overall prevalence is below about 15% (Figure 6), and then decreases sharply and essentially linearly, reaching 0 when the overall prevalence is 100%. When an outcome has an overall prevalence of 50%, our suggested maximum plausible CV is 0.58 (Table 1).

The bounds for σ_L and ICC$_L$ implied by our maximum entropy considerations are shown in Figure 7 and Table 1. It is important to note that, unlike on the probability scale, these quantities do not fall away to zero as they approach 0% and 100%, because σ_L is unbounded above and may be specified independently of μ.

Empirical data

In an interesting investigation, Gulliford et al.[5] quantified the empirical relationship between overall prevalence μ and ICC that was observed in healthcare settings with a large number of clusters and a large number of subjects in each cluster. After swapping the prevalence μ for $1-\mu$ when $\mu > 0.5$, they found linear associations of $\log$ ICC with $\log \mu$ in two datasets of ICCs they compiled. Using the symmetry of ICC about $\mu = 0.5$, the relationship

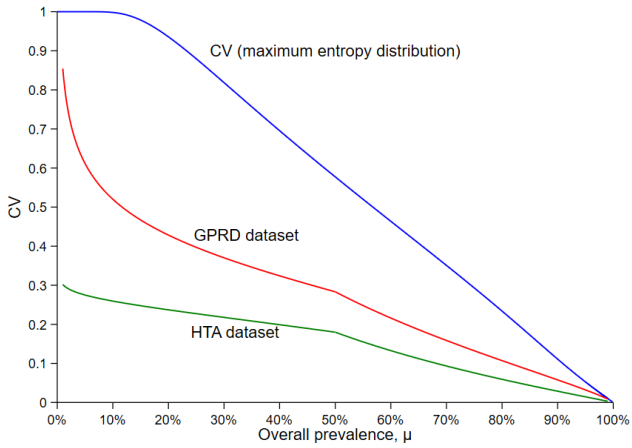

Figure 6. Bounds for the Coefficient of Variation (CV) describing the between-cluster variation in prevalence on the probability scale (assuming the maximum entropy distribution describes the maximum plausible amount of variation).

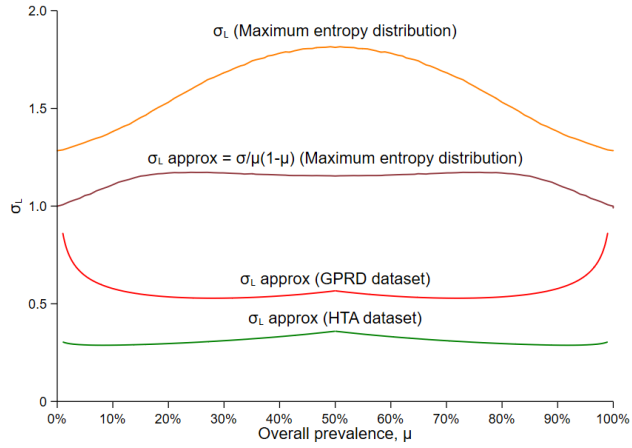

Figure 7. Bounds for the standard deviation describing the between-cluster variation in prevalence on the log odds scale (assuming the maximum entropy distribution describes the maximum plausible amount of variation). (Constructed following simulation of data.)

between ICC and CV ($\text{ICC} = \text{CV}^2 \times \mu/(1-\mu)$) and the aforementioned approximate relationship between σ and σ_L, we have drawn their associations on Figure 5, Figure 6 and Figure 7, with μ ranging from 0.01 to 0.99.

Their estimated regression lines can be seen to fall below our bounds for all $0.01 < \mu < 0.99$. Using the dataset arising from their analysis of the General Practice Research Database (GPRD), average ICCs were 0.008, 0.032 and 0.080 for prevalences of 0.01, 0.1 and 0.5, respectively; corresponding ICCs of 0.001, 0.007 and 0.032 were found for their second dataset concerning outcomes in community and health services settings from a Health Technology Assessment (HTA) review.

The corresponding associations between overall prevalence μ and CV in Figure 6 provide rare, hard-to-find empirical evidence that CV tends to decline as prevalence increases from 1% to 50%[10]. That CV tends to decline as prevalence increases from 50% to 100% is clear from the definition of CV: the SD tends to decrease as the mean increases.

The associations between overall prevalence μ and $\sigma/(\mu(1-\mu))$ in Figure 7 suggest that σ_L, and its approximation, have only a little dependence on prevalence. Our bounds for σ_L, and its approximation, also have only a little dependence on prevalence. Since $\sigma/(\mu(1-\mu))$ is very nearly constant and equal to one, another plausible approximate upper bound for the standard deviation σ is given by $\text{Var}(y_{ij}) = \mu(1-\mu)$.

Recall that our approximation to σ_L is valid only for small σ_L, and so the large differences in Figure 7 between exact and approximate versions of σ_L are not unexpected.

Discussion

Despite appearing in the vast majority of sample size and power calculations of cluster randomised trials, many trialists find the intra-cluster correlation coefficient to be unintuitive in the setting of binary outcomes. Some researchers will be surprised to learn just how much variation is implied by an ICC as low as 0.04. In general, it is hard to decide what is a large amount of variation and what is not. The ICC is typically expressed in terms of variances, not standard deviations, which may explain why more attention is given to variances despite standard deviations being easier to interpret. We suggest that understanding is much improved through visualising the distribution of true cluster prevalences or at least calculating their standard deviation, which is more readily interpretable than the dimensionless ICC—the latter cannot be interpreted independently of the overall prevalence. In principle, this excercise in understanding can and should be done for each arm of a parallel-group cluster randomised trial[7]. It is, perhaps, worth repeating the general advice to think in terms of standard deviations: their units are likely to be better grounded in reality, and in this case are event probabilities.

Outside the trials setting, this paper may be helpful to researchers seeking to understand variation and clustering in any study with a binary outcome. In observational studies where a random intercepts logistic regression model is used, it may be helpful to use the Murray and Murray[15] approximation to calculate the standard deviation σ from σ_L, or to visualise the distribution of p_i (assuming either a beta distribution or a logistic-normal distribution), for one or more fixed prevalences (as in Figure 1).

Appealing to maximum entropy theory, we have proposed bounds on several measures for what between-cluster variation in outcomes is plausible (Table1). This led to the following rule of thumb: that plausible ICCs and standard deviations of true cluster prevalences are both bounded above by the overall prevalence, its complement, and one third. A qualitatively similar bound for the ICC ("ICCs over 0.35 are unlikely for binary outcomes, and for extreme prevalences ICCs may be even smaller") was suggested by Eldrige and Kerry[2] based on a beta distribution being unimodal.

Our rule of thumb may help researchers planning a cluster randomised trial to not specify implausibly high values of the ICC. Implausibly high values of the ICC can be suggested by analysis of previous data (this includes pilot studies and very many trials, even large trials), for at least two reasons. Firstly, ICCs are often estimated very imprecisely, with large standard errors[21]. Low precision in estimating ICCs can be

seen whenever the number of clusters is small, so even in large trials our bounds may provide a helpful reality check. Secondly, ICC_L can be confused with ICC[22], and while the difference between the two may be small when $\mu = 0.5$, for very small (or very large) μ, ICC_L will be several times larger than the ICC, as we and others have shown[6].

Confusing ICC_L for ICC is a particularly easy trap for Stata users to fall into. Stata's simple "estat icc" command calculates ICC_L following mixed effects logistic regression. If a Stata user looks hard in the software's extensive documentation, they may find mention of another definition of ICC for binary data (on the probit scale after use of the "meprobit" command), but no mention of the usual ICC on the probability scale. In fact, this is true even for the required ICC input to the "power twoproportions, cluster" command, which experienced trialists will realise must be the ICC on the probability scale. However, this is not made explicit in the documentation. We also note that the default value taken by this ICC is 0.5, irrespective of the mean prevalence μ, and far beyond our suggested bounds even if $\mu \approx 50\%$.

Our bounds may also help to inform the choice of prior distributions in Bayesian analyses. For example, Turner et al.[23] employed a uniform prior distribution on $[0, 1]$ for the ICC "to represent lack of prior knowledge", but a narrower range may be warranted, particularly if prior information indicates an overall prevalence near zero or one.

A recent example of an anticipated ICC that was unrealistically high is a stepped wedge trial of a cleaning intervention to reduce the rates of healthcare-associated infections in hospitals[24]. The power calculation assumed binary outcome data with an overall prevalence of combined infection (per occupied bed-day) of $\mu = 0.0015$. It specified a within-hospital correlation in infection of 0.3 (far higher than our bound), the source of which is unclear (Adrian Barnett, personal communication). The actual trial data (during the pre-intervention phase) showed an overall prevalence of $\mu = 0.0004$, ICC $= 0.0001$ estimated using one-way ANOVA,[6,25], and $\text{ICC}_L = 0.1$. These three represent very different ICC values, even if stepped wedge trials are relatively insensitive to variations in ICC[26].

Our bound for the ICC is broadly consistent with the findings of Gulliford et al.[5], although when prevalence is 0.5% the average ICC from their model on one of their datasets (GPRD) is 0.005, the same as our bound. The average ICC from their model on the HTA dataset (where the smallest overall prevalance was 0.003% rather than 0.2% (GPRD)) is 0.0005, comfortably below our bound. Gulliford et al.[5] point out "[t]he distribution of cluster-specific proportions [...] may vary according to the nature of the outcome measure, the characteristics of clusters and individual subjects, and the context of a study", and Eldridge and Kerry[2] review patterns in ICCs, such as their apparent dependence on prevalence, and the fact that process outcomes tend to have larger ICCs than clinical outcomes. Almost half of the ICCs in the GPRD dataset arose from studies of the proportion of GP consultations resulting in a prescription of antibiotics (including penicillins, nonpenicillins or penicillins and nonpenicillins combined), for each of fifteen acute conditions. Many other ICCs were derived from studies of proportions of patients in a general practice prescribed each of fifteen classes of drugs.

Our ICC bound (which tends to 0 as the mean prevalence decreases) agrees qualitatively with all of this prior empirical and theoretical research. However, the maximum entropy distribution itself often seems too extreme to be described as "plausible". For example, a uniform distribution (with $\mu = 0.5$) seems implausible to us as a distribution of cluster prevalences. We suggest that values above or close to our bound should be treated with caution, and the reasoning behind this choice should be double-checked. We emphasise, though, that ICCs beyond our bound are of course possible: if patients in different clusters receive very different care because of clinical training or practice that is highly variable internationally, a bimodal distribution might occur and even ICCs approaching 1 may arise.

Our bound for the coefficient of variation (CV $= 1$ for low prevalences) is somewhat higher than that of Hayes et al.[8], who say that "experience from field trials suggest that the coefficient of variation is often ≤ 0.25". A coefficient of variation of 0.25 can arise in a setting where the mean prevalence is expected to be 30%, but prevalence in villages could easily vary between 15% and 45% according to experts[8]. Assuming a roughly normal distribution, this could suggest $\sigma = 0.075$ (so that 95% of villages have a prevalence within 2 standard deviations of the mean), hence CV $= 0.075/0.3 = 0.25$. A similar crude heuristic argument, this time restricting 95% of cluster prevalences to take non-negative values (e.g. 0% to 60%), leads Hayes et al.[8] to suggest coefficient of variation "seldom exceeds 0.5".

The quantity σ_L can be considered transportable within a cluster randomised trial if a treatment acts multiplicatively on the *odds* of an adverse event[7]. Our bound for σ_L (1.28 to 1.81 depending on prevalence) fits nicely with the range (up to 0.9) often seen in cluster randomised trials[6]. This bound has less dependence on prevalence than the bounds for both ICC and CV. The approximation $\sigma_L \approx \sigma/(\mu(1-\mu))$ has a bound that has less dependence on prevalence still. Between-cluster variation quantified in this way on previous data has little dependence on prevalence[5].

Dedication

This paper is dedicated to the memory of Dr Dan Lunn, our undergraduate mathematics and statistics tutor at Worcester College, Oxford. His depth of insight made him unorthodox—we recall his prodigal proof that the sample mean and variance are independent—while his inexhaustible supply of anecdotes kept us laughing all the way to the Examination Schools. Dan made statistics both meaningful and merry, and we both owe him a huge debt of gratitude.

References

1. Glorioso TJ, Grunwald GK, Ho PM et al. Reference effect measures for quantifying, comparing and visualizing variation from random and fixed effects in non-normal multilevel models, with applications to site variation in medical procedure use and outcomes. *BMC Medical Research Methodology* 2018; 18(1): 74. DOI:10.1186/s12874-018-0517-7. URL https://doi.org/10.1186/s12874-018-0517-7.
2. Eldridge S and Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. John Wiley

& Sons, 2012. ISBN 978-1-119-96672-2. Google-Books-ID: UZnEbtweiDQC.

3. Kahan BC. Accounting for centre-effects in multicentre trials with a binary outcome – when, why, and how? *BMC Medical Research Methodology* 2014; 14(1): 20. DOI:10.1186/1471-2288-14-20. URL https://doi.org/10.1186/1471-2288-14-20.
4. Rutterford C, Copas A and Eldridge S. Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology* 2015; 44(3): 1051–1067. DOI:10.1093/ije/dyv113. URL https://academic.oup.com/ije/article/44/3/1051/632956.
5. Gulliford MC, Adams G, Ukoumunne OC et al. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology* 2005; 58(3): 246–251. DOI:10.1016/j.jclinepi.2004.08.012. URL http://www.sciencedirect.com/science/article/pii/S0895435604002938.
6. Eldridge SM, Ukoumunne OC and Carlin JB. The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. *International Statistical Review* 2009; 77(3): 378–394. DOI:10.1111/j.1751-5823.2009.00092.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2009.00092.x.
7. Thomson A, Hayes R and Cousens S. Measures of between-cluster variability in cluster randomized trials with binary outcomes. *Statistics in Medicine* 2009; 28(12): 1739–1751. DOI:10.1002/sim.3582. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3582.
8. Hayes RJ, Moulton LH and Moulton LH. *Cluster Randomised Trials*. Chapman and Hall/CRC, 2017. ISBN 978-1-315-37028-6. DOI:10.4324/9781315370286. URL https://www.taylorfrancis.com/books/9781315370286.
9. Crespi CM, Wong WK and Wu S. A new dependence parameter approach to improve the design of cluster randomized trials with binary outcomes. *Clinical Trials* 2011; 8(6): 687–698. DOI:10.1177/1740774511423851. URL https://doi.org/10.1177/1740774511423851.
10. Chatfield MD and Farewell DM. Letter to the Editor: Is the R coefficient of interest in cluster randomized trials with a binary outcome? *Statistical Methods in Medical Research* 2020; 29(6): 1763–1764. DOI:10.1177/0962280220912783. URL https://doi.org/10.1177/0962280220912783. Publisher: SAGE Publications Ltd STM.
11. Merlo J, Chaix B, Ohlsson H et al. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology & Community Health* 2006; 60(4): 290–297. DOI:10.1136/jech.2004.029454. URL https://jech.bmj.com/content/60/4/290. Publisher: BMJ Publishing Group Ltd Section: Continuing professional education.
12. Breslow NE and Clayton DG. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 1993; 88(421): 9–25. DOI:10.2307/2290687. URL www.jstor.org/stable/2290687.
13. Lee EW and Dubin N. Estimation and sample size considerations for clustered binary responses. *Statistics in Medicine* 1994; 13(12): 1241–1252. DOI:10.1002/sim.4780131206. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780131206. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780131206.
14. StataCorp. *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC, 2019.
15. Murray DM. *Design and Analysis of Group-randomized Trials*. Oxford University Press, 1998. ISBN 978-0-19-512036-3. Google-Books-ID: 9ERYAgAAQBAJ.
16. Abramowitz M and Stegun IA. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Courier Corporation, 1965. ISBN 978-0-486-61272-0. Google-Books-ID: MtU8uP7XMvoC.
17. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 1994; 6(4): 284–290. DOI:10.1037/1040-3590.6.4.284.
18. Bhatia R and Davis C. A Better Bound on the Variance. *The American Mathematical Monthly* 2000; 107(4): 353–357. DOI:10.2307/2589180. URL https://www.jstor.org/stable/2589180. Publisher: Mathematical Association of America.
19. Jaynes ET. Information Theory and Statistical Mechanics. *Physical Review* 1957; 106(4): 620–630. DOI:10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620.
20. Conrad K. Probability Distributions and Maximum Entropy. Technical report, University of Connecticut, 2014. URL https://kconrad.math.uconn.edu/blurbs/analysis/entropypost.pdf.
21. Cook JA, Bruckner T, MacLennan GS et al. Clustering in surgical trials - database of intracluster correlations. *Trials* 2012; 13(1): 2. DOI:10.1186/1745-6215-13-2. URL https://doi.org/10.1186/1745-6215-13-2.
22. Martin J, Girling A, Nirantharakumar K et al. Intra-cluster and inter-period correlation coefficients for cross-sectional cluster randomised controlled trials for type-2 diabetes in UK primary care. *Trials* 2016; 17(1): 402. DOI:10.1186/s13063-016-1532-9. URL https://doi.org/10.1186/s13063-016-1532-9.
23. Turner RM, Prevost AT and Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Statistics in Medicine* 2004; 23(8): 1195–1214. DOI:10.1002/sim.1721. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1721.
24. Mitchell BG, Hall L, White N et al. An environmental cleaning bundle and health-care-associated infections in hospitals (REACH): a multicentre, randomised trial. *The Lancet Infectious Diseases* 2019; 19(4): 410–418. DOI:10.1016/S1473-3099(18)30714-X. URL http://www.sciencedirect.com/science/article/pii/S147330991830714X.
25. Ridout MS, Demétrio CGB and Firth D. Estimating Intraclass Correlation for Binary Data. *Biometrics* 1999; 55(1): 137–148. DOI:10.1111/j.0006-341X.1999.00137.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.1999.00137.x.
26. Baio G, Copas A, Ambler G et al. Sample size calculation for a stepped wedge trial. *Trials* 2015; 16(1): 354. DOI:10.1186/s13063-015-0840-9. URL https://doi.org/10.1186/s13063-015-0840-9.

Appendix: entropy, and maximum entropy distributions

We have used entropy in this paper to characterise uncertainty in probability distributions, and compute it for various examples in Figures 1–4. This appendix provides a little more detail about entropy and maximum entropy distributions.

Informally, entropy measures the average information content in a probability distribution, where a degenerate distribution carries no information and where entropy increases as the distribution becomes more evenly spread out across its support.

More formally (and slightly more broadly), entropy is a quantification of the average ability to discriminate between a given probability distribution and some arbitrary reference distribution. Since the most powerful numerical summary for discrimating between two candidate models is their likelihood ratio R, entropy is typically defined as $-E(\log R)$, with expectation being taken with respect to the distribution of interest rather than the reference distribution.

For continuous distributions on the unit interval such as the majority of those dealt with in the present paper, the uniform distribution is often chosen as the point of reference, and a distribution with probability density function $f(x)$ has a so-called differential entropy of

$$-\int_0^1 f(x)\log f(x)\,\mathrm{d}x.$$

By construction, the uniform distribution itself has an entropy value of 0. The beta distribution $\mathrm{Beta}(\alpha,\beta)$ has differential entropy

$$\begin{aligned}&\log B(\alpha,\beta)-(\alpha-1)\psi(\alpha)\\&\qquad-(\beta-1)\psi(\beta)+(\alpha+\beta-2)\psi(\alpha+\beta),\end{aligned}$$

where B is the beta function and ψ the digamma function. In the absence of a closed form expression for the entropy of a logistic normal distribution, numerical integration was used to compute the entropy in Figure 1. Differential entropy can be negative, and discrete distributions may be assigned a notional differential entropy of $-\infty$.

Conrad[20] illustrates how to derive maximum entropy distributions on a given support and with particular constraints. Of particular interest to us is Conrad's Theorem 5.1, which states that "[t]he continuous probability density function on the interval $[a,b]$ with mean μ that maximizes entropy among all such densities (on $[a,b]$ with mean μ) is a truncated exponential density". When $[a,b]=[0,1]$, this has the form

$$f(x)=\frac{c\exp(cx)}{\exp(c)-1}$$

for the unique c satisfying the equation $\mu=1-1/c+1/\{\exp(c)-1\}$. The variance of this maximum entropy distribution may be computed (e.g. via integration by parts), and turns out to be

$$\frac{1}{c^2}+\frac{1}{2-2\cosh c}.$$

More straightforwardly, its differential entropy is

$$\log\{\exp|c|-1\}-\log|c|-|c|\mu$$

for $c\neq 0$, and is 0 if $c=0$ (the uniform distribution).