On the Genetic Interpretation of Disease Data

OPEN ACCESS Freely available online

PLoS one

On the Genetic Interpretation of Disease Data

Stephen C. Bishop*, John A. Woolliams

The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, Midlothian, United Kingdom

Abstract

Background: The understanding of host genetic variation in disease resistance increasingly requires the use of field data to obtain sufficient numbers of phenotypes. We introduce concepts necessary for a genetic interpretation of field disease data, for diseases caused by microparasites such as bacteria or viruses. Our focus is on variance component estimation and we introduce epidemiological concepts to quantitative genetics.

Methodology/Principal Findings: We have derived simple deterministic formulae to predict the impacts of incomplete exposure to infection, or imperfect diagnostic test sensitivity and specificity on heritabilities for disease resistance. We show that these factors all reduce the estimable heritabilities. The impacts of incomplete exposure depend on disease prevalence but are relatively linear with the exposure probability. For prevalences less than 0.5, imperfect diagnostic test sensitivity results in a small underestimation of heritability, whereas imperfect specificity leads to a much greater underestimation, with the impact increasing as prevalence declines. These impacts are reversed for prevalences greater than 0.5. Incomplete data recording in which infected or diseased individuals are not observed, e.g. data recording for too short a period, has impacts analogous to imperfect sensitivity.

Conclusions/Significance: These results help to explain the often low disease resistance heritabilities observed under field conditions. They also demonstrate that incomplete exposure to infection, or suboptimal diagnoses, are not fatal flaws for demonstrating host genetic differences in resistance, they merely reduce the power of datasets. Lastly, they provide a tool for inferring the true extent of genetic variation in disease resistance given knowledge of the disease biology.

Citation: Bishop SC, Woolliams JA (2010) On the Genetic Interpretation of Disease Data. PLoS ONE 5(1): e8940. doi:10.1371/journal.pone.0008940

Editor: Syed A. Aziz, Health Canada, Canada

Received September 8, 2009; **Accepted** November 3, 2009; **Published** January 28, 2010

Funding: The authors were funded by the Biotechnology and Biological Sciences Research Council (BBSRC), via The Roslin Institute's Institute Strategic Programme Grant (ISPG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: S.C.B. has on one occasion served as an academic editor (08-PONE-RA-04516) for a Plos One paper.

* E-mail: Stephen.Bishop@roslin.ed.ac.uk

Introduction

Genetic variation in host resistance to infectious disease is ubiquitous [1,2,3]. The increasing realization of this phenomenon has led to disease biology becoming a major focus of ecology and population or quantitative genetic research for human and animal geneticists alike. Further, the ready availability of dense single nucleotide polymorphism arrays (i.e. SNP chips) has given rise to hitherto unforeseen opportunities to dissect this between-host variation and identify possible genes contributing to this variation using genome wide association studies [4]. This, coupled with more traditional quantitative genetic variance-partitioning approaches [5], enables detailed descriptions of genetic aspects of disease resistance and the identification of individuals with extreme (high or low) risk of infection or disease [6]. Such techniques can be applied equally to human, natural animal populations or farmed livestock.

To have the requisite power to meaningfully quantify genetic variation or perform a genome scan using a dense SNP chip it is necessary to have datasets comprising observations on several thousands of individuals [e.g. 7]. For studies of infectious diseases this usually necessitates utilizing field data because challenge experiments of a sufficient scale will not be possible, possibly excepting studies with aquacultural species [e.g. 8]. For example, in the livestock context, data may be captured from a population undergoing an epidemic such as bovine tuberculosis [9], or from an endemic disease such as mastitis [see 10], where the herd-level prevalence is largely predictable. However, such field data is very 'noisy': diagnosis of infection or disease may be imprecise; it can be difficult to determine when infection of an individual occurred; and it is often unclear whether or not apparently healthy individuals have been exposed to the infection. These factors will add environmental noise to the epidemiological data.

Issues such as exposure and diagnostic test sensitivity or specificity are fundamental concepts to epidemiologists when studying the spread of disease in a population [11], yet their intrinsic importance is currently ignored in quantitative genetic theory [5]. Quantifying and accounting for the impact of environmental factors is an integral part of identifying and measuring true host genetic variation in resistance to the disease under study. Consequently, there is an unrecognised risk of biases in genetic parameter estimates and lost opportunities for identifying individuals with extreme genetic risk. This paper proposes advances in quantitative genetic theory using concepts borrowed from epidemiology and provides predictive equations for the impact of epidemiological factors on heritability estimation. The theory is developed specifically for microparasitic infections, such as those caused by viruses or bacteria.

Analysis

General Framework

Consider a generic microparasitic disease in which individuals may move between infection states as illustrated in Figure 1. Upon exposure to infection a *susceptible* (i.e. not yet infected) individual may become infected and *infectious*, after which it may either recover or die. For simplicity, the states of *diseased* and *infectious* are considered equivalent in this study. The term *susceptible* does not indicate an individual's liability to infection; rather, it denotes that it is not immunologically resistant and can become infected. If *susceptible* individuals are replenished, either through loss of immunity of *recovered* individuals or through immigration of new individuals, then an endemic equilibrium may be reached in which the expected disease prevalence is constant. Otherwise, under assumptions of homogeneous random mixing the number of infected individuals will ultimately go to zero, and the epidemic will die out with the expected proportion of individuals ever infected during the course of the epidemic (I^*) satisfying the equation $I^* = 1 - e^{(-R_0 I^*)}$ [12], where R_0 is the basic reproductive ratio of the disease. Therefore, assuming no disease-independent mortality, the expected proportion of susceptible individuals remaining in the population at the completion of the epidemic is $1 - I^*$.

Inferences about host genetic resistance are generally made by comparing *diseased* and *healthy* individuals. The *diseased* category will include infected and/or dead individuals, and the *healthy* category will include *susceptible* individuals, i.e. not yet infected, and possibly *recovered* individuals. In more complex models, individuals with latent infection that have yet to display detectable signs of infection may also be included in the *healthy* category. Heritabilities are determined by estimating to what degree the expected genetic relationships predict the classification of individuals into *healthy* and *diseased*, whereas individual SNP associations are inferred from departures of SNP allele frequencies from their expectations within the two categories. The genetic associations uncovered by such analyses will indicate host genetic variation in 'disease resistance', where the term 'disease resistance' is used generically to cover any of the processes shown Figure 1 that may influence the probability of an individual being diagnosed as *diseased*.

Several sources of uncertainty in field disease data can be identified from Figure 1. Firstly, for an individual to move from the *susceptible* to the *latently infected* or *infectious* category, it is necessary for it to be exposed to infection. A lack of exposure simply means that individuals do not have the opportunity to express their genotype for resistance, with potentially highly susceptible individuals being classified as *healthy*. In a group of individuals one might quantify exposure by ε, the probability that an individual is exposed to infection. Secondly, the diagnostic test used to classify individuals as *healthy* or *diseased* may be imperfect, with individuals misclassified. Specificity (S_p) measures the probability that a healthy individual is classified as *healthy* by the diagnostic test, whereas sensitivity (S_e) measures the probability that a diseased individual is classified as *diseased* by the test [11]. Thirdly, it is apparent from Figure 1 that an epidemic is a dynamic process. When data are collected over any time period which is less than the duration of the epidemic, the outcomes may differ from the outcomes that would have been obtained if the data were to have been collected over the entire epidemic, again through misclassifications.

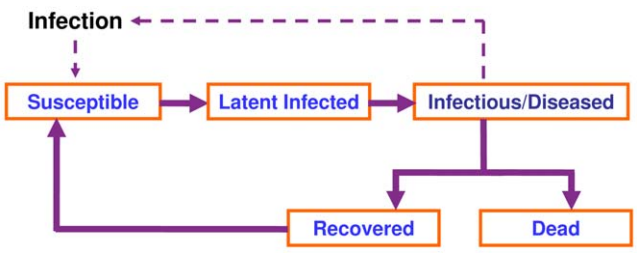

Figure 1. Model for transmission of bacterial or viral infections.
doi:10.1371/journal.pone.0008940.g001

These three phenomena whilst distinct are not independent, i.e. they are interrelated outcomes of the properties of the epidemic. For example, exposure probabilities may depend on the duration of data recording, with the probability of exposure increasing with time. However, for development of quantitative theory, their impacts are described and interpreted separately. The impacts of incomplete exposure and diagnostic test sensitivity and specificity can be explored independent of the epidemic dynamics, and hence are termed static disease properties. The impacts of time-dependent measurements require dynamic disease epidemic models.

Static Disease Properties

(a) Incomplete Exposure to Infection. When there is incomplete exposure to infection the observed prevalence, the fraction of the whole population that is identified as *diseased* is a function of two factors: (i) the proportion of individuals that have been exposed to the pathogen (ε), and (ii) the virtual prevalence (p), which is defined as the proportion of individuals that have been exposed to the pathogen that become infected. Assuming that exposure is random and independent of host genotype, then the observed prevalence is εp. Of the $1-\varepsilon p$ proportion of individuals that are *healthy*, $\varepsilon(1-p)$ are exposed and apparently resistant, whilst $(1-\varepsilon)$ have not yet been exposed and have not expressed any genotype related to 'disease resistance'. The phenotypic variance of observed 'disease resistance' is given by the binomial variance $\varepsilon p(1-\varepsilon p)$.

Firstly, consider the epidemic among the exposed, with virtual prevalence p. Suppose that on the underlying liability scale the heritability is h^2 for true disease resistance, i.e. resistance following actual exposure, and the total liability has variance 1. Then using the linear approximation often used in the genetic analyses of binary traits [13], the genetic variance expressed on the binomial 0/1 scale is given by $\phi(x_p)^2 h^2$ where x_p is the truncation point of the Normal distribution corresponding to upper-tail probability p, and $\phi(x_p)$ is the corresponding Normal density. Now consider the case of incomplete exposure and let D'_u and D'_w be the observed states (either *healthy*, 0, or *diseased*, 1) of individuals u and w with numerator relationship a_{uw}, and let Z be an indicator trait with $Z=1$ if both u and w are exposed and $Z=0$, otherwise. Assuming exposure is independent of the numerator relationship then $\mathrm{cov}(D'_u, D'_w | Z=1) = a_{uw}\phi(x_p)^2 h^2$ and $\mathrm{cov}(D'_u, D'_w | Z=0) = 0$, so $\mathrm{cov}(D'_u, D'_w | Z) = a_{uw}\phi(x_p)^2 h^2 Z$; when $Z=0$ the covariance is not expressed since at least one individual is not exposed, and there is only one outcome for that individual, $D'=0$. Then using the general formula for unconditional covariances: $\mathrm{cov}(D'_u, D'_w) = E[\mathrm{cov}(D'_u, D'_w | Z)] + \mathrm{cov}(E[D'_u|Z], E[D'_w|Z])$ and noting (i) the latter term is 0, and (ii) $E(Z)=\varepsilon^2$ the probability of both being exposed, the result emerges: $\mathrm{cov}(D'_u, D'_w) = a_{uw}\varepsilon^2\phi(x_p)^2 h^2$.

Therefore on the 0/1 scale the true heritability of disease resistance is $\phi(x_p)^2 h^2 p^{-1}(1-p)^{-1}$ whilst the observed heritability is $\varepsilon\phi(x_p)^2 h^2 p^{-1}(1-\varepsilon p)^{-1}$. This differs by a factor $\varepsilon(1-p)/(1-\varepsilon p)$. This will always be ≤ 1 since both $\varepsilon \leq 1$ and $(1-p)/(1-\varepsilon p) \leq 1$. Furthermore, this biased heritability is transformed back to the liability scale as kh^2, where $k = \varepsilon^2\phi(x_p)^2/\phi(x_{\varepsilon p})^2$. The

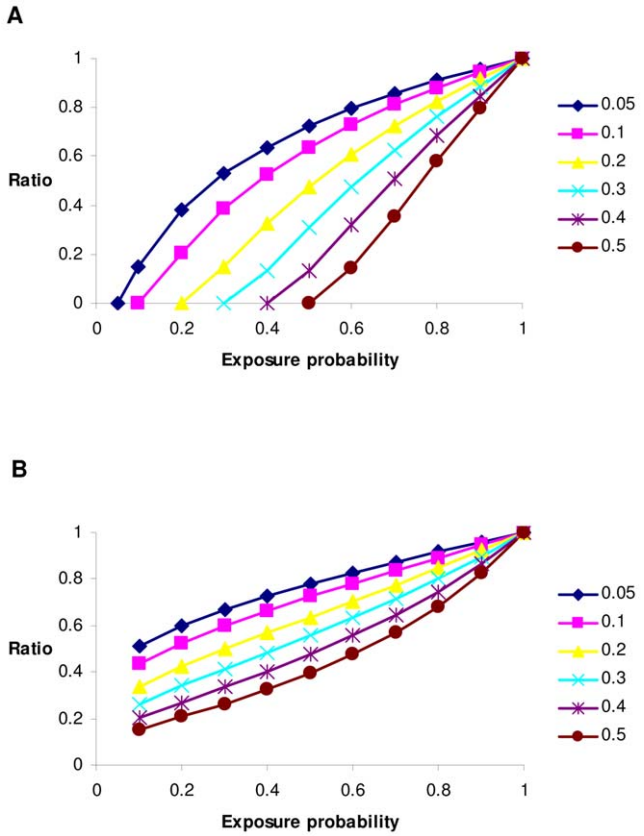

Figure 2. Ratio of estimated to true heritability on the liability scale for incomplete exposure. Results are shown for (A) differing observed prevalences or (B) differing virtual prevalences.
doi:10.1371/journal.pone.0008940.g002

bias on the liability scale is less than that on the observed scale since the reduced prevalence that is observed due to incomplete exposure leads to a greater scaling of the observed heritability back to the liability scale. For small εp, the under-prediction on the 0/1 scale is close to a linear function of ε. The bias is greater if p is moderate or large.

Impacts of differing exposure probabilities and differing virtual prevalences are illustrated in Figures 2a and 2b where observed and virtual prevalences are varied, respectively. In both cases the exposure probability has a close to linear impact on the bias parameter. The bias is more severe when considering the relationship as a function of observed prevalence, because when the exposure probability drops towards the observed prevalence, it implies the *healthy* population is dominated by individuals that have not been exposed to infection.

(b) Incomplete Sensitivity and Specificity of Diagnostic Tests. Individuals will be classified into *healthy* and *diseased* categories by means of a diagnostic test for the disease of interest. Fundamental to any diagnostic test are the concepts of specificity and sensitivity. As described above, specificity (S_p) is the probability that a truly *healthy* individual is classified by the diagnostic test as *healthy* and sensitivity (S_e) is the probability that a truly *diseased* individual is classified by the diagnostic test as *diseased*. The implications of sensitivity and specificity on the proportions of individuals diagnosed as healthy or diseased are shown in Table 1. The true prevalence is given as p, and the prevalence observed from the diagnostic test is p'.

Insight into the column margins can be gained by observing that (S_p+S_e-1) is the regression coefficient of the classification based upon the diagnostic test on the true state where disease is scored 1 and healthy 0. The regression line is $D'=p'+(S_p+S_e-1)D$. As above, let D_u and D_w be the true classification of individuals u and w with numerator relationship a_{uw}. The impact of imperfect S_e and S_p on estimates of heritability can be deduced assuming that the classification errors are independent for u and w, and unrelated to D_u or D_w. The covariance between the observed classification D'_u and D'_w can be obtained from $\mathrm{cov}(D'_u, D'_w)=E[\mathrm{cov}(D'_u, D'_w|D_u, D_w)]+\mathrm{cov}(E[D'_u|D_u, D_w], E[D'_w|D_u, D_w])$. The first of these terms is identically zero given the assumption made. The second term is then the covariance of the terms in Table 2, which can be derived from the regression line above. This gives the result $\mathrm{cov}(D'_u,D'_w)=(S_p+S_e-1)^2\mathrm{cov}(D_u,D_w)$. It then follows directly that if u and w have a genetic covariance of $a_{uw}h^2$ on the liability scale then $\mathrm{cov}(D_u,D_w)=a_{uw}\phi(x_p)^2h^2$ and $\mathrm{cov}(D'_u,D'_w)=a_{uw}\phi(x_p)^2h^2(S_p+S_e-1)^2$ with observed prevalence p'. Thus, the observed heritability on the 0/1 scale is $h'^2=\phi(x_p)^2h^2(S_p+S_e-1)^2p'^{-1}(1-p')^{-1}$ and when transformed back to the liability scale it is $\phi(x_p)^2h^2(S_p+S_e-1)^2\phi(x_{p'})^{-2}$.

Impacts of various specificities and sensitivities on estimated heritability values are illustrated in Figures 3a and 3b, where only sensitivity and specificity, respectively, are varied and 3c, in which they are varied jointly. For all prevalences, imperfect sensitivity and specificity both result in underestimated heritabilities on the liability scale. However the impact of poor specificities is much greater, for true prevalence less than 0.5. The reason for this difference is that when decreasing S_e, the term (S_p+S_e-1) decreases, and the observed prevalence p' decreases also, so although $(S_p+S_e-1)^2<1$, this is partially compensated by $\phi(x_p)^2\phi(x_{p'})^{-2}>1$. In contrast, when S_p decreases, the observed prevalence p' increases, and so both $(S_p+S_e-1)^2<1$ and $\phi(x_p)^2\phi(x_{p'})^{-2}<1$. When both sensitivity and specificity are imperfect, then liability-scale heritabilities are considerably underestimated. This is likely to be the case in many practical situations, indicating that true genetic variation in disease resistance is likely to be much greater than indicated by analyses of field data.

Table 1. Proportions of individuals classified as Healthy or Diseased, as a function of Specificity (S_p) or Sensitivity (S_e).

		Classification by diagnostic test:		
		Healthy	*Diseased*	*Total*
True State:	*Healthy*	$(1-p)S_p$	$(1-p)(1-S_p)$	$1-p$
	Diseased	$p(1-S_e)$	pS_e	p
	Total	$1-p'=S_p-p(S_p+S_e-1)$	$p'=(1-S_p)+p(S_p+S_e-1)$	

doi:10.1371/journal.pone.0008940.t001

Table 2. Covariance expectations between animals with different disease classification status.

D_u	D_w	Probability	$E[D_u'\|D_u,D_w]$	$E[D_w'\|D_u,D_w]$
1	1	$p^2+\mathrm{cov}(D_u,D_w)$	S_e	S_e
1	0	$p(1-p)-\mathrm{cov}(D_u,D_w)$	S_e	$1-S_p$
0	1	$(1-p)p-\mathrm{cov}(D_u,D_w)$	$1-S_p$	S_e
0	0	$(1-p)^2+\mathrm{cov}(D_u,D_w)$	$1-S_p$	$1-S_p$

doi:10.1371/journal.pone.0008940.t002

Dynamic Disease Properties

The principle of dynamic epidemic models is that individuals move between infection state categories, as shown in Figure 1. At different points during the epidemic it may be different individuals that are observably *diseased*, and the efficiency with which all potentially *diseased* individuals (I^*) are observed as *diseased* depends on the duration of the data collection period in relation to the dynamics of the epidemic. In most data recording scenarios lasting for a time period Δt, i.e. temporally incomplete data recording, only a proportion of individuals ever transiting through the *infectious/diseased* categories will be observed. Let the total number of individuals observed to be *infectious/diseased* in the interval t to $t+\Delta t$ be defined as $I(t,\Delta t)$ therefore the proportion of all individuals ever *diseased* that are observed is $I(t,\Delta t)/I*$. This is analogous to imperfect diagnostic test sensitivity. Therefore, the impact of temporally incomplete data recording on estimated heritabilities is the same as for imperfect sensitivity.

As an illustration of the impact of dynamic disease properties, consider a simple SIR model [12]. Let $S(u)$ and $I(u)$ be the instantaneous number of susceptible and infectious animals at time u, and β be the transmission coefficient for the disease. For a recording period starting at time $u=t$, and lasting for time period Δt, then $I(t,\Delta t)=I(t)+\int_t^{t+\Delta t}\beta S(u)I(u)du$. Therefore, the ratio $I(t,\Delta t)/I*$ will depend not only on the duration of the recording period Δt, but also when recording commenced in relation to the epidemic. This ratio will be termed the 'epidemic sensitivity'.

As an illustration, consider an SIR model with parameters $\beta=0.00015$, $\gamma=0.1$, where γ is the recovery rate, $R_0=1.5$ and hence $I^*=0.59$. For this parameterization, and starting with one infected individual, it will take approximately 180 days for 95% of all individuals potentially infected during an epidemic to become *diseased*. It is assumed that recording starts when the disease prevalence reaches 5% and that the diagnostic test is perfect, i.e. sensitivity and specificity are both unity. Two scenarios are considered, (i) where only *infectious/diseased* individuals are observed, and (ii) where *recovered/removed*, e.g. dead, individuals are also observed. Plotted in Figure 4 are the proportions of individuals ever *diseased* during the course of the epidemic that are observed during the observation period, i.e. the epidemic sensitivity $I(t,\Delta t)/I*$. Observations taken only at one time point will result in a low epidemic sensitivity, hence underestimated heritabilities, and observations taken at different start points will also vary. If both *diseased* and *recovered/removed* individuals are observable, then the epidemic sensitivity becomes high with an extended observation period, since individuals that are infected and recover or removed prior to recording are also observed. However, if *recovered* individuals are not observable, i.e. they are healthy and no longer show any symptoms or clinical signs, then the epidemic sensitivity remains low and heritabilities remain underestimated.

Discussion

This paper has provided a framework to assist in the interpretation of field disease data, with extensions to quantitative genetics theory being presented to account for the effects of various forms of environmental noise on genetic parameters for disease resistance. The factors considered, viz. incomplete recording, incomplete exposure, imperfect sensitivity and specificity of diagnosis are all typical of the non-genetic influences encountered with field disease data. We demonstrate in this paper that the likely impacts of these factors on genetic parameters for disease resistance are largely predictable, provided ball park figures can be obtained for specificity, sensitivity or exposure probabilities. In summary, estimable heritabilities are biased downwards by each of these factors. Conversely, the presence of detectable genetic variation in field disease data implies that the true heritability for disease resistance, were it to be measured under ideal circumstances, is likely to be much higher.

A further significance of the theory presented in this paper is that it can reconcile our observation that whilst traits describing immune responses to infection are often highly heritable, the disease outcomes that these traits influence tend to be lowly

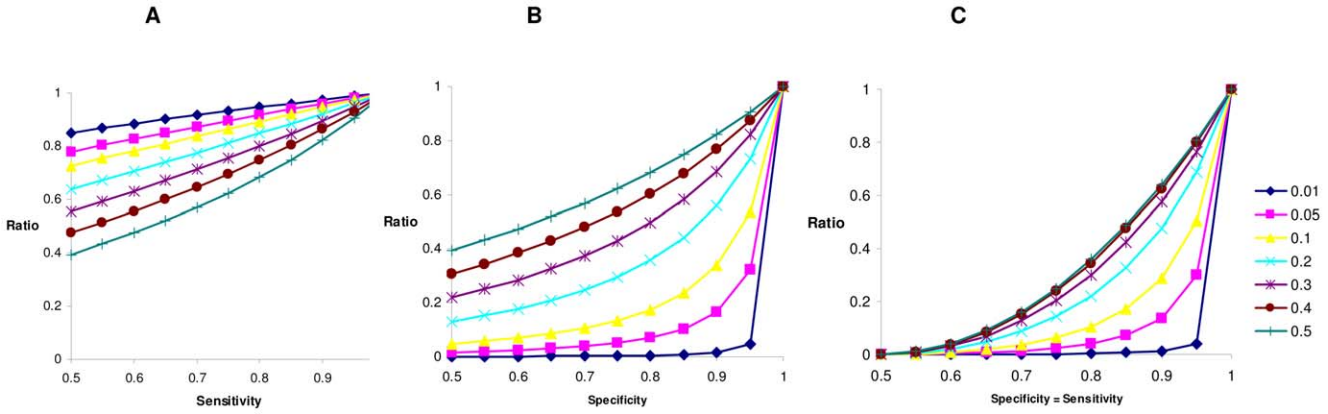

Figure 3. Ratio of estimated to true heritability on the liability scale for differing true prevalences. Results are shown for (A) incomplete sensitivity, where specificity = 1, (B) incomplete specificity, where sensitivity = 1 or (C) for incomplete specificity and sensitivity, where the two parameters equal.
doi:10.1371/journal.pone.0008940.g003

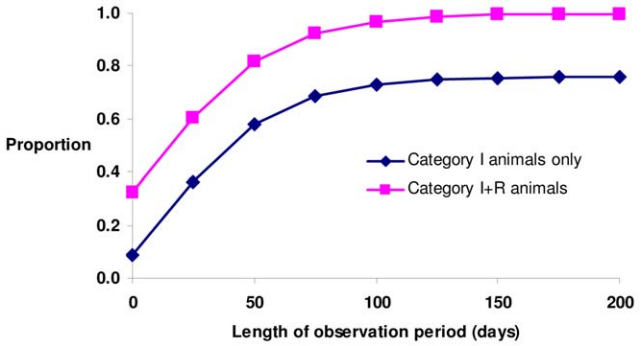

Figure 4. An example of the proportion of individuals recorded as *infectious/diseased* relative to those ever *infectious/diseased* during an SIR epidemic, as a function of recording period. Two cases are shown, with only I individuals observable or with both I and R observable. In this example, recording is triggered when prevalence reaches 5%. Parameters in this model are: $\beta = 0.00015$, $\gamma = 0.1$ and $R_0 = 1.5$.
doi:10.1371/journal.pone.0008940.g004

heritable. This is best illustrated from extensive datasets collected in farmed livestock. For example, components of innate and adaptive immunity are often moderately to highly heritable in commercial pig populations [14,15], whereas the heritability of observable disease in such animals is low to moderate at best [16,17]. Whilst true presence or absence of disease, given exposure to infection, will be largely a function of the immune response, we have demonstrated that the actual prevalence of disease and the estimable genetic variation between animals will be influenced by variable exposure and the sensitivity of diagnosis. Similarly, in commercial dairy cattle, many studies have demonstrated that the occurrence of clinical mastitis invariably has a heritability less than 0.1 [10], whereas underlying immune responses to infection display heritabilities which though variable are often high [e.g. 18].

Published field data are available which supports the concepts developed in this paper. For example, predicted impacts of exposure to infection on estimable heritabilities may be inferred from data recently published on resistance to infectious pancreatic necrosis (IPN), a viral disease affecting farmed salmon. Heritabilities for IPN-related survival of salmon located in seawater localities containing the IPN virus were estimated and presented for seven independent cohorts of fish [19]. Of these seven cohorts, five fulfilled criteria of comprising populations unselected for IPN resistance and having heritability values consistent with the observed prevalence, i.e. heritabilities transformed to the liability scale [13] remained within the parameter space. For these five cohorts, the observed prevalences were 0.10, 0.12, 0.14, 0.19 and 0.30 and the corresponding heritabilities on the observed (0,1) scale were 0.11, 0.20, 0.16, 0.28 and 0.56, respectively, showing the expected strong relationship between prevalence and heritability for this scale. In principle, transformation to the liability scale should remove the relationship between prevalence and heritability, but the values obtained (0.32, 0.53, 0.39, 0.59 and 0.97) continue to show a significant linear relationship with prevalence. Because these five cohorts may be regarded as subpopulations sampled at random in relation to IPN resistance from the same overall population, it may be hypothesized that the differences in prevalence simply reflect differences in exposure rates. Relative exposure probabilities in each cohort may therefore be estimated as the ratio of observed prevalence to that seen in the cohort with the highest prevalence. Estimating exposure probability in this way, and using the above theory to rescale the heritability for liability, resulted in the heritabilities displayed in Figure 5, along with the regression of these heritabilities on observed prevalence. The strong linear relationship between prevalence and the heritability of liability to IPN disappeared when differences in relative exposure probabilities were hypothesized and the induced biases were removed. Furthermore it suggests that the heritability is large and important.

The heritability of resistance to bovine tuberculosis in dairy cattle provides an example of the potential impact of diagnostic test sensitivity and specificity on observable genetic variation. A recent publication provided convincing evidence of moderate genetic variation in tuberculosis resistance in dairy cattle, with an average heritability of liability of 0.12 in a dataset with a prevalence of 0.10 [9]; further, this paper speculated that imperfect sensitivity and specificity may have resulted in an underestimation of the true heritability. At this prevalence, imperfect specificity has a large impact on the estimated heritability, however the specificity of this diagnostic test is likely to be high. Sensitivity may be lower, possibly closer to 0.8 [20]. Exploring scenarios for specificities of 0.98 or 0.99, and sensitivities varying between 0.7 and 0.9, leads to the conclusion that the observed heritability is possibly underestimated by 20 to 40%. Therefore, the true heritability in this population is likely to be in the range 0.15 to 0.20.

Sometimes, particularly in an animal breeding context, an indicator trait is used to describe the impact of infection or disease upon an individual, for example somatic cell count in the milk of lactating ruminants with mastitis [10]. Hence, the measurements comprise a mixture distribution, i.e. those taken on both *healthy* and *diseased* individuals. These data may be analysed ignoring the fact that some individuals are *healthy* and others *diseased*, however this potentially leads to misleading results if the statistical properties of the trait (variance, heritability, etc) differ between the two subpopulations, or if the biological interpretation of the indicator trait differs between the two subpopulations. For example, dairy cattle breeders may wish to select on somatic cell count to reduce the incidence of mastitis, but they may not wish to alter mean somatic cell count in *healthy* cows [10]. Ideally, the data could be split into *healthy* and *diseased* subpopulations, and analysed separately. Various methods based on the properties of the data distribution have been proposed to achieve this [21]; alternatively an independent diagnostic of infection may be used, such as the presence of mastitis-causing microorganisms in the milk. Whatever approach is used, the concepts of diagnostic test accuracy still apply and biases may occur if these are ignored. For example the

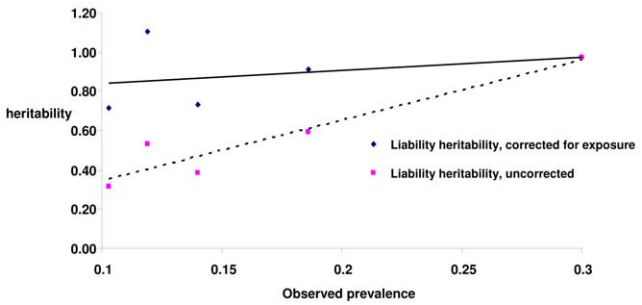

Figure 5. Heritabilities for liability to death from infectious pancreatic necrosis in five cohorts of Atlantic salmon, before and after correction for inferred relative exposure levels. The data are from Guy *et al.* 2009 [19]. Shown are heritability values and linear regression trend lines.
doi:10.1371/journal.pone.0008940.g005

true difference in the indicator trait between the subpopulations will be underestimated for imperfect sensitivity or specificity, as animals will be misclassified.

We now determine the impact of imperfect sensitivity and specificity on the properties of indicator traits such as somatic cell count. If H_i and D_i are indicator trait observations in truly *healthy* or *diseased* subpopulations, and H'_i and D'_i are indicator trait observations in an imperfectly classified population in which the observed prevalence is p', then the estimated true difference between diseased and healthy individuals $(\Delta=\mu_D-\mu_H)$ is, after simplification, $\Delta=(\mu_{D'}-\mu_{H'})\{(S_p+S_e-1)p(1-p)/[p'(1-p')]\}^{-1}$. For plausible S_p and S_e values, Δ is always greater than $(\mu_{D'}-\mu_{H'})$. Similarly, properties of the variances of the observed subpopulations can be estimated from the properties of mixture distributions, and they contain an upwards bias proportional to Δ^2. We have applied these concepts to mastitis in sheep (Riggio, Bishop and coworkers, unpublished data), using a dataset where diagnoses were available for the mastitis infection status of every ewe on every occasion that somatic cell count measurements were taken. These data demonstrated that specificity and sensitivity of diagnosis must have been high, as poor values would have led to implausible Δ values. Given high but plausible specificity and sensitivity (>0.9), inferred genetic correlations between the indicator trait measured in *healthy* and *diseased* animals were moderate (ca. 0.6) and insensitive to small changes in either parameter.

The theory presented in this paper does contain a number of simplifying assumptions, most notably that exposure probability or diagnostic test sensitivity and specificity are independent of host genotype. These assumptions may sometimes be violated. As an example, related individuals may be more likely to be co-exposed to infection, e.g. family members in the same household or animals in the same litter, and this potentially introduces a bias into heritability estimation. An issue may also arise with diagnostic tests in which animal immune responses are measured, such as skin test measurements used to infer exposure to bovine tuberculosis [20]. If aspects of these immune responses are genetic in origin, as seems plausible, this may impact on diagnostic test sensitivity. We have yet to fully explore the impact of these factors on expected genetic parameter values.

Many disease genetic studies now bypass the step of estimating variance components to quantify genetic variation and move directly to SNP association studies, unfortunately ignoring the design information that may give an objective assessment of the plausibility of both the design and the outcomes of the study. Nevertheless, the principles and consequences of noisy field data for the estimation of SNP effects are analogous to those for variance component estimation. For example, with incomplete exposure a fraction $(1-\varepsilon)/(1-\varepsilon p)$ of individuals that are *healthy* have not been exposed and hence do not contribute information. Therefore, the effective size of the control population is smaller by this proportion. Furthermore, with imperfect sensitivity and specificity, there is a reduction in the estimable SNP effect size by (S_p+S_e-1) due to the regression coefficient of the diagnostic classification on the true state, with a consequent reduction in the experimental power for detecting SNP associations.

In summary, we believe that the results presented in this paper add clarity to the interpretation of field disease data, and reduce the risk that incorrect inferences are made regarding the extent of genetic variation. We have considered the different aspects of field data separately, but the underlying theory is clear and the potential exists to combine the different factors to match specific scenarios. We suggest that published estimates of heritabilities for resistance to microparasitic diseases, corresponding SNP effects and study design should be re-appraised given knowledge of the disease biology, i.e. likely exposure to infection, properties of the diagnostic tests and duration of data recording.

Acknowledgments

The study was prompted in part by consideration of bovine TB epidemiology in a project funded by Defra, UK.

Author Contributions

Conceived and designed the experiments: SCB JAW. Performed the experiments: SCB JAW. Analyzed the data: SCB. Contributed reagents/materials/analysis tools: JAW. Wrote the paper: SCB.

References

1. Segal S, Hill AVS (2003) Genetic susceptibility to infectious disease. Trends in Microbiology 11: 445–448.
2. Trammell RA, Toth LA (2008) Genetic susceptibility and resistance to influenza infection and disease in humans and mice. Expert Review of Molecular Diagnostics 8: 515–529.
3. Bishop SC (2005) Disease resistance: Genetics. In: Pond WG, Bell AW, eds. Encyclopedia of animal science. New York: Marcel Dekker, Inc.. pp 288–290.
4. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 6: 95–108.
5. Lynch M, Walsh B (1998) Genetics and Analysis of Quantitative Traits. SunderlandMassachusetts: Simauer Associates. 980 p.
6. Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE 3(10): e3395. doi:10.1371/journal.pone.0003395.
7. Tenesa A, Farrington SM, Prendergast JGD, Porteous ME, Walker M, et al. (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nature Genetics 40: 631–637.
8. Kjoglum S, Henryon M, Aasmundstad T, Korsgaard I (2008) Selective breeding can increase resistance of Atlantic salmon to furunculosis, infectious salmon anaemia and infectious pancreatic necrosis. Aquaculture Research 39: 498–505.
9. Bermingham ML, More SJ, Good M, Cromie AR, Higgins IM, et al. (2009) Genetics of tuberculosis in Irish Holstein-Friesian dairy herds. Journal of Dairy Science 92: 3447–3456.
10. Rupp R, Boichard D (2003) Genetics of resistance to mastitis in dairy cattle. Veterinary Research 34: 671–688.
11. Noordhuizen JPTM, Frankera K, van der Hoofd CM, Graat EAM (1997) Application of Quantitative Methods in Veterinary Epidemiology. Wageningen: Wageningen Pers. 445 p.
12. Anderson RM, May RM (1992) Infectious Diseases of Humans. Dynamics and Control. Oxford: Oxford University Press. 757 p.
13. Robertson A, Lerner IM (1949) The heritability of all-or-none traits: viability of poultry. Genetics 34: 395–411.
14. Henryon M, Heegaard PMH, Nielsen J, Berg P, Juul-Madsen HR (2006) Immunological traits have the potential to improve selection of pigs for resistance to clinical and subclinical disease. Animal Science 82: 597–606.
15. Clapperton M, Glass EJ, Bishop SC (2008) Pig peripheral blood mononuclear leucocyte sub-sets are heritable and genetically correlated with performance. Animal 2: 1575–1584.
16. Henryon M, Berg P, Jensen J, Andersen S (2001) Genetic variation for resistance to clinical and subclinical diseases exists in growing pigs. Animal Science 73: 375–387.
17. Henryon M, Berg P, Christensen G, Jensen J, Lund MS, et al. (2003) Visual assessment of post-mortem lesions exhibits little additive genetic variation in growing pigs. Livestock Production Science 83: 121–130.
18. Detilleux JC, Koehler KJ, Freeman AE, Kehrli ME, Kelley DH (1994) Immunological parameters of periparturient Holstein cattle - genetic-variation. Journal of Dairy Science 77: 2640–2650.
19. Guy DR, Bishop SC, Woolliams JA, Brotherstone S (2009) Genetic parameters for resistance to Infectious Pancreatic Necrosis in pedigreed Atlantic salmon populations using a reduced animal model. Aquaculture 290: 229–235.
20. de la Rua-Domenech R, Goodchild AT, Vordermeier HM, Hewinson RG, Christiansen KH, et al. (2006) Ante mortem diagnosis of tuberculosis in cattle: A review of the tuberculin tests, gamma-interferon assay and other ancillary diagnostic techniques. Research in Veterinary Science 81: 190–210.
21. Gianola D, Heringstad B, Odegaard J (2006) On the quantitative genetics of mixture characters. Genetics 173: 2247–2255.