# Effect of changes in testing parameters on the cost-effectiveness of pooled testing to classify infection status of animals in a herd

*Locksley L. McV. Messam[1,2], Joshua M. O'Brien[3], Sharon K. Hietala[4], Ian A. Gardner[5]

[1]St. Georges University, Department of Public Health and Preventive Medicine, School of Medicine, P.O. Box 7, True Blue, St. Georges, Grenada, West Indies.

[2]St. Georges University, Office of the Dean, School of Veterinary Medicine, P.O. Box 7, True Blue, St. George's, Grenada, West Indies.

[3]Center for Animal Disease Modeling and Surveillance, University of California, Davis, California, 95616, USA

[4]California Animal Health and Food Safety Laboratory System, Davis CA, 95616, USA

[5]Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California, Davis, California, 95616, USA

*Corresponding author. Tel.: +1 473 444 4175 ext. 2623; Fax: +1 473 444 1219.
*E-mail address*: lmessam@sgu.edu (L. L. McV. Messam)

## Abstract

Monte Carlo simulation was used to determine optimal fecal pool sizes for identification of all *Mycobacterium avium subsp. paratuberculosis* (MAP)-infected cows in a dairy herd. Two pooling protocols were compared: a halving protocol involving a single retest of negative pools followed by halving of positive pools and a simple protocol involving single retest of negative pools but no halving of positive pools. In the simulations, the distributions of number of tests required to classify all individuals in an infected herd were generated for various combinations of prevalence (0.01, 0.05 and 0.1), herd size (300, 1000 and 3000), pool size (5, 10, 20 and 50) and test sensitivity (0.5 to 0.9). Test specificity was fixed at 1.0 because fecal culture for MAP yields no or rare false-positive results. Optimal performance was determined primarily on the basis of a comparison of the distributions of numbers of tests needed to detect MAP-infected cows using the Mann-Whitney U test statistic. Optimal pool size was independent of both herd size and test characteristics, regardless of protocol. When sensitivity was the same for each pool size, pool sizes of 20 and 10 performed best for both protocols for prevalences of 0.01 and 0.1, respectively, while for prevalences of 0.05, pool sizes of 10 and 20 were optimal for the simple and halving protocols, respectively. When sensitivity decreased with increasing pool size, the results changed for prevalences of 0.05 and 0.1 with pool sizes of 50 being optimal especially at a prevalence of 0.1. Overall, the halving protocol was more cost effective than the simple protocol especially at higher prevalences. For detection of MAP using fecal culture, we recommend use of the halving protocol and pool sizes of 10 or 20 when the prevalence is suspected to range from 0.01 to 0.1 and there is no expected loss of sensitivity with increasing pool size. If loss in sensitivity is expected and the prevalence is thought to be between 0.05 and 0.1, the halving protocol and a pool size of 50 is recommended. Our findings are broadly applicable to other infectious diseases under comparable testing conditions.



Keywords

Cost-effectiveness, pooled testing, *Mycobacterium avium subsp. paratuberculosis,* retesting

## 1. Introduction

For studies of animal disease, pooled or group testing has 3 main applications: prevalence estimation (Cowling et al., 1999), herd or group classification (Christensen and Gardner, 2000) and identification of individuals with a particular characteristic, usually infection. In the last decade, pooled samples have been increasingly used for infectious disease testing (Kalis et al., 2000; Borel et al., 2004; Jouy et al., 2005; Kennedy et al., 2006; Singer et al., 2006; Brinkhof et al., 2007; Kennedy et al., 2008). Methodological research for animal diseases has focused on comparing methods for individual-level prevalence estimation from pooled samples (Cowling et al., 1999; Evers and Nauta, 2001; Williams and Moffitt, 2005; Toribio and Sergeant, 2007), evaluating the effect of prevalence and pool size on diagnostic test characteristics (Maherchandani et al., 2004; Sanderson et al., 2005; Munoz-Zanzi et al., 2006; Rovira et al., 2007; van Schaik et al., 2007), identifying economically optimal protocols and pool sizes for screening (Munoz-Zanzi et al., 2000; van Schaik et al., 2003) and estimating the costs of testing for particular pooling scenarios (Munoz-Zanzi et al., 2000; van Schaik et al., 2003; van Schaik et al., 2007). Overall, published evidence indicates that when infection is rare, substantial savings can be achieved when samples are initially tested in pools rather than individually tested (Abel et al., 1999; Kalis et al., 2000; van Schaik et al., 2003).

Nevertheless, basic questions persist despite methodological advances; including what is the best pool size to use in order to cost effectively and accurately identify all infected individuals when given a particular protocol and prevalence of infected animals in a herd. While a few studies have addressed this topic (Gupta and Malina, 1999; van Schaik et al., 2003), results have not been easy to apply under field conditions. Guidelines for choosing optimal pool size are often reported in a statistically complex manner and presented in terms of the pool size that results in the lowest expected number of tests (or lowest expected cost of testing) per individual (Dorfman, 1943; Sterrett, 1957; Patel, 1962; Graff and Roeloffs, 1972; Gwa et al., 1992; Gastwirth and Johnson, 1994; Gupta and Malina, 1999; Munoz-Zanzi et al., 2000; van Schaik et al., 2003). One problem with this approach is that expected values are summaries of distributions of

numbers of tests (or total costs) given a pre-specified pool size and protocol, and provide no information on the likelihood that an investigator will exceed a particular cost threshold for that particular combination of protocol and pool size. In addition, an expected value gives neither an indication of the shape of the distribution of number of tests it summarizes nor the extent of overlap with a distribution resulting from another pool size. Consequently, a diagnostician wishing to use pooled testing does not have all the key elements on which to base an informed decision regarding pool size. For a given prevalence, protocol, diagnostic test, and sample size, investigators would benefit from information about the distribution of the total number of tests (or total costs) for a variety of possible pool sizes. In this way, comparisons could be made between the distributions of numbers of tests for different scenarios to determine which scenarios result in larger numbers of tests and to evaluate the probabilities of incurring costs within particular ranges (determined by budgetary constraints) given various pool sizes. Such an investigator would also have information on the range of costs that might be incurred (e.g. 2.5 and 97.5 percentiles of the distribution) for several protocols, where available, and could assess the risk associated with a particular choice of pool size.

There is a need for practical decision-making guidelines for pool size determination in herds infected with *Mycobacterium avium subsp. paratuberculosis* (MAP). The subclinical nature of Johne's disease necessitates testing to identify most infected animals and commonly-used diagnostic tests have low sensitivity (e.g. approximately 30% for serum and milk ELISAs and 60% for fecal culture)(Nielsen and Toft, 2008). The cost-effective identification and elimination of animals shedding MAP is used as an adjunct to management practices to control transmission. Investigators of Johne's disease, and others faced with similar scenarios, will benefit from the use of information on the distribution of total numbers of tests required to identify all infected animals in a herd in the context of pooling protocols which assure high predictive values. These will help investigators to make more realistic budgetary predictions and thereby facilitate evidence-based decision making.

The aim of the present study was to identify optimal pool sizes for various combinations of testing protocol, test sensitivity, test specificity and prevalence for identification of all infected cows in a dairy herd known to be MAP infected. The optimal

pool size was primarily determined by a comparison of the distribution of the number of tests used to classify all infected animals in the herd for various pool sizes. Given a particular testing protocol, prevalence, sensitivity and specificity, the pool size that generated a distribution with stochastically lowest values for the number of tests was considered optimal.

## 2. Materials and Methods

We used Monte Carlo simulation to compare two pooled testing protocols because the distribution of numbers of tests associated with a given protocol, pool size, diagnostic test, sample size and sampling scheme can be calculated and compared prior to study implementation. Under random sampling, this information would match what occurs under field conditions.

### *2.1 Prevalence*

For the purpose of the study, we assumed that an accurate estimate of MAP infection prevalence $(p)$ in the herd(s) was available. This would apply, for instance to herds participating in the U.S. Johne's disease Demonstration Herd Project (Groenendaal and Wolf, 2008). For many of these herds, the true (individual-level) MAP infection prevalence $(p)$ is < 0.1. Consequently for simulations, populations were generated with values of $p$ = 0.01, 0.05 and 0.1. For each prevalence, herds of size $N = 300$, 1000 and 3000 were created.

### *2.2 Diagnostic test characteristics*

Specificity $(Sp)$ was fixed at 1.0 while sensitivity $(Se)$ was increased from 0.5 to 0.9, in increments of 0.1. Scenarios reflecting imperfect Se but perfect Sp are justified on the basis that the most frequently used test for identification of MAP infected individuals in

both pooled and individual testing is fecal culture on solid or in liquid media (Kalis et al., 2000; van Schaik et al., 2003; Collins et al., 2006). Fecal culture has variable Se depending on whether the animal is a low, moderate or heavy shedder and near perfect specificity $(Sp \approx 0.999)$. Because Se is imperfect, pools of size $k$ which initially test negative, are retested once to ensure results with high negative predictive value $(NPV \geq 0.9)$. Perfect positive predictive value $(PPV = 1.0)$ is assured if $Sp = 1.0$.

*2.3 Pool sizes and protocols*

The effect of using pool sizes of 5, 10, 20 and 50 was compared for two different protocols under the assumption that $Se$ was independent of pool size.

*2.3.1 Modified Simple Pooling (MSP) Protocol.*

This method involves random allocation of all individuals in a herd of size $N$ and of infection prevalence $p$, into pools of size $k$ (Fig.1). In the first stage, all $n_p (= N/k)$ pools are tested using a test with $Se$ varying from 0.5 to 0.9 and $Sp = 1$. In the second stage, each pool that initially tests negative is then retested once. In the absence of scientific data, we assumed that the sensitivity is reduced by a half each time a sample is retested to account for the likely conditional dependence in test sensitivity induced by repeated testing of the same infected sample (Greiner and Gardner, 2000). If the retested pool tests negative, all of the constituent individual samples are considered uninfected. If a pool tests positive during the retesting procedure, it is assumed that the test-positive pool contains at least 1 infected sample. In the third stage, individual samples in test-positive pools are tested with $Se$ from 0.5 to 0.9 (i.e. the initial sensitivity used before the retesting) and $Sp = 1$. Individual samples that test positive are considered infected and those that test negative are considered uninfected (see Fig. 1). This is a modification of the method initially suggested by Dorfman (Dorfman, 1943) and used by some veterinary researchers (Letellier et al., 2005; Singer et al., 2006). The retesting component is designed to increase the *NPV* of the testing procedures by increasing the overall Se of the procedure when compared to the same protocol without retesting. For this protocol, the

expected number of tests $\left(n_{t(1)MSP}\right)$ needed to classify all individuals in one pool of size $k$ is:

$$E\left(n_{t(1)MSP}\right)=\underbrace{(k+1)\,p_p Se}_{A}+\underbrace{(k+2)\left[\left(1-p_p\right)+p_p\left(1-Se\right)\right]\left[p_p\left(0.5\left(Se\right)\right)\right]}_{B}$$
$$+\underbrace{2\left[\left(1-p_p\right)+p_p\left(1-Se\right)\right]\left[\left(1-p_p\right)+p_p\left(1-0.5\left(Se\right)\right)\right]}_{C}$$

where A is the contribution to the total number of tests if the pool initially tests positive, B is the contribution if the pool tests positive upon retesting after having initially tested negative once and C is the contribution if the pool tests negative upon retesting after having initially tested negative (see appendix I and II for derivation). Here $p_p$ is the prevalence of positive pools. The expected number of tests for the whole herd = $n_p \times E\left(n_{t(1)MSP}\right)$.

*2.3.2 Modified Halving (MH) Protocol.*

This protocol differs from the MSP protocol, in that individuals in all pools that test positive (those that initially test positive plus those that test positive upon retesting) are randomly allocated to two pools of size $k/2$ for even $k$ (or $1+\left(k-1\right)/2$ and $\left(k-1\right)/2$ for odd $k$) and retested (Fig. 1). Pools that test negative at this stage are considered negative, while those testing positive are subjected to individual testing of component samples. For the MH protocol, the expected number of tests $\left(n_{t(1)MH}\right)$ needed to classify all individuals in one pool of initial size $k$ is:

$$E\left(n_{t(1)MH}\right)=\underbrace{\left(p_p Se\right)\left[\sum_{i=1}^{q}\left(\left(1+m_q\right)\frac{\binom{N_p}{q}\binom{N-N_p}{k-q}}{\binom{N}{k}}\right)\right]}_{A'}$$

$$+\underbrace{\left[\left(1-p_p\right)+p_p\left(1-Se\right)\right]\left[p_p\left(0.5\left(Se\right)\right)\right]\left[\sum_{i=1}^{q}\left(\left(2+m_q\right)\frac{\binom{N_p}{q}\binom{N-N_p}{k-q}}{\binom{N}{k}}\right)\right]}_{B'}$$

$$+\underbrace{2\left[\left(1-p_p\right)+p_p\left(1-Se\right)\right]\left[\left(1-p_p\right)+p_p\left(1-0.5\left(Se\right)\right)\right]}_{C'}$$

Here $A'$ is the contribution to the total number of tests if the pool initially tests positive and is then halved, $B'$ is the contribution if the pool tests positive and is halved after having initially tested negative and been retested once, $C'$ is the contribution if the pool tests negative after having initially tested negative and been retested, $m_q$= the expected number of subsequent tests after halving a pool of initial size k containing q infected samples and $\frac{\binom{N_p}{q}\binom{N-N_p}{k-q}}{\binom{N}{k}}=\Pr\left(\text{exactly q infected samples in pool of size k}\right)$. (see appendices I and III). Note that $A'$ and $B'$ depend on the number of infected samples contained in a test-positive pool before it is halved as previously described (Munoz-Zanzi et al., 2000). The expected number of tests for the whole herd is $n_p \times E\left(n_{t(1)MH}\right)$.

*2.4 Simulations*

Monte Carlo simulations were used to mimic the pooled testing process for herds of various sizes with each run producing a numerical value reflecting the number of tests required to classify all animals in the herd. The simulations were executed in R (R, Development Core Team 2005) and the code is available on request. For the simulations,

$k$ individuals were randomly selected without replacement from a population of $N$ individuals to create $n_p$ pools each of size $k$ using the Mersenne Twister pseudo-random number generator (Matsumoto and Nishimura, 1998). We performed 5000 simulations because we speculated that in the context of testing of a dairy, if a particular total number of tests does not occur at least once in 5000 trials it would not be of practical importance for herd sizes of 300, 1000 and 3000. Results from simulations closely approximated analytic results for MSP. This was also the case for MH for $k = 5$ and $10$ and $p = 0.01$ where the probability of $\geq 3$ infected samples being in a pool was negligible. This provided assurance that the code was accurate and would work for combinations of larger pool sizes and prevalences in which the probabilities of $\geq 3$ infected samples being in a pool was non-negligible and analytically tedious to calculate.

*2.4.1 Performance measures.*

The optimal pool size for a given combination of herd size, prevalence, test characteristics and testing protocol was determined on the basis of a comparison of 4 performance measures described below in decreasing order of importance:

a. The Mann-Whitney U test (Mann and Whitney, 1947; Altman, 1991) was used to determine the probability that a randomly chosen value from distributions of numbers of tests generated by pool sizes 5, 20 and 50, respectively, were stochastically greater than a randomly chosen value from a distribution of numbers of tests generated by a pool size of 10. Given two distributions $X$ and Y with elements $x_1, x_2, ..., x_i$ and $y_1, y_2, ..., y_j$, respectively, the Mann-Whitney U statistic provides the number of pairs for which $x_i > y_j$. A pool size of 10 was chosen for comparison as this is a commonly-used pool size for fecal culture testing for MAP. Probabilities were calculated by dividing the respective U statistics by 25,000,000 (the number of possible pairs when one element is drawn from two sets containing 5000 elements each)(Altman, 1991). The lower the probability the more economically advantageous the pool size was considered to be. Values of 0 and 1 represent distributions with no overlap and always have

values stochastically lower and greater, respectively, than values from distributions generated by a pool size of 10.

b. The probability of savings from pooled testing being at least 75 % of the cost of individual testing for the equivalent scenario i.e. $\Pr\left(C_{T(P)} < (1/4)C_{T(I)}\right)$ given $C_{T(P)} = n_p^* C_p + n_{t(n_p)} C_t$ and $C_{T(I)} = n_{t(N)} C_t$ where $C_{T(P)}$ = the total cost of pooled testing, $C_{T(I)}$ = the total cost of individual testing, $C_p$ = the unit cost of pool creation, $C_t$ = the unit cost of testing, $n_{t(n_p)}$ = the total number of tests performed on pooled samples, $n_{t(N)}$ = the number of tests performed during a protocol using individual testing and $n_p^*$ = the number of pools created during the procedure. For MSP, $n_p^* = n_p = N/k$ , while for MH, $n_p^* = n_p + 2n_p^+$, where $n_p^+$ = the number of pools that test positive during MH and from which $2n_p^+$ new pools are to be formed. We estimated $C_p$ and $C_t$ to be \$1 and \$20, respectively. The higher the probability the more economic the pool size was considered to be. Overhead costs and costs associated with sample collection were excluded because these were assumed to be independent of the strategies applied.

c. The probability of savings from pooled testing being at least 50 % of the cost of individual testing for the equivalent scenario i.e. $\Pr\left(C_{T(P)} < (1/2)C_{T(I)}\right)$

d. The 2.5, 50th and 97.5 percentiles of the distributions of percent savings $\left(\left(\left(C_{T(I)} - C_{T(P)}\right)/C_{T(I)}\right)\times 100\right)$ . The relative performance of pools of different sizes was determined by comparing the percentiles (2.5, 50 and 97.5) of their distributions (in that order of importance). Pool sizes with higher 2.5 and 50th percentiles were considered more economical and given a higher rank even if they had lower 97.5 percentiles. We contend that since the goal of pooled testing is to maximize savings, then higher lower percentiles of the distribution of savings should have greater weight than higher upper percentiles.

For a given protocol, pool sizes were first ranked from first to fourth for optimality based on their U statistic and successive performance measures were used in the order presented above (ie. b, c, d) to resolve ties when necessary. Initially, ranking was performed for both the MSP and MH Protocols separately and then comparisons were made between protocols using criteria b, c and d. Minimum negative predictive values (NPV) for each testing scenario (distribution of numbers of tests) were also calculated.

*2.4.2 Sensitivity analysis*.

To test the robustness of our results to changes in test characteristics for each scenario, we repeated our simulations with the following modifications. First, pools which initially tested negative were retested with sensitivities equal to one-fifth and one-tenth of the initial sensitivity value. This served to evaluate whether rank order was a function of the magnitude of sensitivity correlation between initial and subsequent tests. Second, test sensitivities were varied as a function of pool size as follows: sensitivities for pool sizes of 50 were 80 % of the sensitivities for pool size of 20, sensitivities for pool sizes of 20 were 90% of sensitivities for pool sizes of 10 and sensitivities for pool sizes of 10 were 90 % of sensitivities for pool sizes of 5. Both modifications are realistic because the sensitivity correlation might be substantially greater than the value (one-half) we used in the default scenario and for pool sizes ranging from 5 to 50, test sensitivity is only likely to remain invariant with respect to pool size for herds in which there are highly infectious animals (high fecal shedders of MAP).

## 3. Results

For both protocols, the direction of changes in median percentage savings and the probability of savings being at least 50% and 75% of the cost of individual testing as a function of changes in prevalence, herd size and test sensitivity were identical and are summarized in Table 1. For given prevalence, sensitivity and test protocol, rank order in terms of optimality of pool sizes did not change with herd size, and hence for brevity, we present results for only a herd size of 1000.

*3.1 Prevalence = 0.01*

For MSP, pool sizes of 20 (Mann Whitney U = 0.08 to 0.22) and 10 (U = 0.50) ranked first and second, respectively, for Se between 0.5 and 0.9 (Table 2). For MH, pool sizes of 20 (U < 0.01) and 50 (U = 0.01 to 0.36) ranked first and second, respectively. Pool sizes of 10 and 5 consistently ranked third and fourth when MH was used regardless of Se, while when MSP was used, pools of 50 ranked higher than pools of 5 for all combinations of herd size and Se. The MH protocol performed better than MSP based on the probability of achieving savings of at least 75% of the cost of individual testing, for each pool size, except a pool size of 5 (where they performed equally well) (Table 3). For a pool size of 5, on the basis of the percentiles of the distributions of savings as a percentage of individual testing costs, MH was more economical than MSP (Se = 0.5 - 0.7) and performed equally at higher sensitivities (Table 4). However, these differences were minor (≤ 1 %). The minimum NPV was 0.99 for both protocols (Table 5). Overall, using the various criteria for assessing optimality, $MH_{20} > MH_{50} > MH_{10} > MSP_{20} > MSP_{10} > MSP_{50} > MH_5 > MSP_5$ (Table 6).

*3.2 Prevalence = 0.05*

For MSP, pool sizes of 10 (U = 0.50) ranked first regardless of Se while pools sizes of 20 (U = 0.80 to 0.99) ranked higher than pools sizes of 5 at Se between 0.5 and 0.7 (U = 0.80 to 0.95 vs. 0.99 to 0.96) and vice versa for Se from 0.8 to 0.9 (U = 0.98 to 0.99 vs. 0.93 to 0.87) (Table 2). However, a pool size of 20 ranked no lower than third using this criterion, while a pool size of 5 was least optimal at lower sensitivities. Pool size of 50 performed least optimally, ranking no greater than third regardless of sensitivity. Both pool sizes of 5 and 20 had equal probabilities of savings being at least 75 % of the cost of individual testing. When we considered the probability of savings being at least 50% of the cost of individual testing (Table 7), compared to a pool size of 5, a pool size of 20 ranked higher at Se between 0.5 and 0.7 but equal for Se between 0.8 and 0.9. For MH, pool sizes of 20 ranked higher than pool sizes of 10 at Se from 0.5 to

0.7 (U = 0.10 to 0.42 vs. 0.50) and vice versa for Se from 0.8 to 0.9 (0.66 to 0.87 vs. 0.50) (Table 2). However, a pool size of 20 had at least as high a probability of achieving savings of at least 75% the cost of individual testing as a pool size of 10 (Table 3). For most Se values, pools of 50 and 5 consistently ranked third and fourth, respectively, based on the U statistic,. MH performed at least as optimally as MSP based on the probability of achieving savings of at least 75% and 50% of the cost of individual testing (Tables 3 and 7, respectively), for each pool size. The minimum NPV for both protocols ranged from 0.95 to 0.98 for Se between 0.5 and 0.9, with MSP being marginally larger (Table 5). Overall, using the various criteria for assessing optimality, $MH_{20} > MH_{10} > MH_{50} > MSP_{10} > MH_5 > MSP_{20} > MSP_{50} > MSP_5$ (Table 6).

*3.3 Prevalence = 0.1*

When MSP was used, pool size of 10 ranked higher than a pool size of 5 at Se between 0.5 and 0.6 (0.50 vs. 0.82 to 0.63) and vice versa for Se between 0.7 and 0.9 (0.5 vs. 0.43 to 0.10) (Table 2). However, a pool size of 10 ranked no lower than second using this criterion while a pool size of 5 was least optimal at Se = 0.5. Pool sizes of 20 and 50 ranked third and fourth, respectively, at most sensitivities (Se from 0.6 to 0.9). Using the MH protocol, a pool size of 20 ranked higher than a pool size of 10 at Se between 0.5 to 0.7 (0.21 to 0.41 vs. 0.5) and vice versa for Se between 0.7 and 0.9 (0.62 to 0.96 vs. 0.5) (Table 2). However, a pool size of 10 had at least as high a probability of achieving savings of at least 75% and 50% the cost of individual testing compared with a pool size of 20 (Tables 3 and 7, respectively). For most sensitivities (Se from 0.6 to 0.9), pool sizes of 50 and 5 ranked third and fourth, respectively. The MH protocol performed at least as well as MSP based on the probability of achieving savings of at least 50% and 75% of the cost of individual testing, for each corresponding pool size, and with the exception of pool size of 5, all MH pool sizes had higher probabilities than MSP pool sizes. On the basis of the percentiles of the distributions of savings as percentages of the cost of individual testing, MH for pool size of 5, ranked higher than all MSP pool sizes (Table 4). The minimum NPV for MSP and MH ranged from 0.91 to 0.97 and 0.90 to 0.96,

respectively, for Se between 0.5 and 0.9 (Table 5). Overall, using the various optimality criteria, $MH_{10} > MH_{20} > MH_{50} > MH_5 > MSP_{10} > MSP_5 > MSP_{20} > MSP_{50}$ (Table 6).

*3.4 Sensitivity analysis*

The use of retesting sensitivities of one-tenth and one-fifth of the initial test sensitivity did not change the rank order of optimality for pool sizes and protocols as compared to when retests were conducted with a sensitivity reduced by one half of the original. At a prevalence of 0.01, the reduction of sensitivity with increase of pool size did not modify the rank order as found when sensitivity was constant for all pool sizes. This however was not the case at prevalences of 0.05 and 0.1 (Table 6) where for the modified halving protocol, the order of optimality was $MH_{50} > MH_{20} > MH_{10}$. At a prevalence of 0.1, among MSP protocols a pool size of 50 performed most optimally.

## 4. Discussion

Our general results indicate that for a herd with MAP prevalence between 0.01 and 0.1, pool sizes of 20 and 10 provide the greatest likelihood and pool sizes of 5, least likelihood of cost savings, respectively, over individual testing. On the other hand, our sensitivity analyses suggest that if the "no loss of sensitivity assumption" is untenable, a pool size of 20 is optimal only at lower prevalences (0.01) and a pool size of 50 has greatest likelihood of savings particularly when the prevalence is 0.1 and when MH is used. This implies that in a herd where highly infectious cows (high or super-shedders of MAP) are present, it is advisable to use a pool size of 20 or 10 depending on the prevalence, while if it is believed that no highly infectious animals are present then a pool size of 50 is advisable. Although no consistent pattern was evident regarding the order of performance for pool sizes for either protocol, several conclusions can be drawn from our results.

First, while the finding that optimum pool size is a function of the true prevalence of infection in the herd is not novel (Dorfman, 1943), the observation that as long as the test sensitivity is the same for all pool sizes, the optimum pool size is relatively unaffected by changes in test sensitivity (Se = 0.5 to 0.9) has not been previously reported. In only 4 scenarios did the ranking of pool sizes change with a change in sensitivity. This only pertained to the number of tests used to classify all individuals and was only evident for pool sizes that had consecutive ranks. The implication of this is that when using either MSP or MH, a Johne's disease investigator need not be concerned with a reevaluation of the choice for optimal pool size due to improvements to the sensitivity of fecal culture unless perfect sensitivity is achieved. The finding that herd size does not affect the relative rankings of pool sizes is in contrast to what has been reported elsewhere (van Schaik et al., 2003). However, in addition to the fact that no retesting was used in that study, the authors presented their results in terms of expected values of costs and thus their results are difficult to compare with ours. Our results suggest that use of a larger pool size solely on the grounds of larger herd size might not be justified.

Second, given constant test sensitivity across pool sizes, larger pool sizes will not necessarily lead to greater savings over individual testing. When sensitivity did not change with pool size, regardless of protocol, pool sizes of 50 performed less favorably in our simulation study than pool sizes of 10 and 20 except in one instance. This result suggests that larger pool sizes will not necessarily decrease the total number of tests required even if detection sensitivity remains unchanged. The apparent contradiction can be explained by the fact that while larger pool sizes have potential to reduce the total amount of individual tests required to classify all individuals, they also increase the probability that all pools will contain an infected sample. Thus there is a trade-off between the potential reduction in the number of tests and the probability of having many pools containing positive samples. The fundamental utility of pooling as a cost-saving procedure depends on the number of positive pools being small. The apparent optimal performance of larger pool sizes when we assumed a loss in sensitivity with increased pool size is attributable to the increased misclassification of infected pools as non-infected pools. We explain this in a following section.

Third, the superiority of the MH over MSP seemed to increase as the prevalence of MAP increased. This is supported by the finding that at a prevalence of 0.1 the least economical MH pool sizes ($MH_5$) performed better than the most economical MSP pool sizes ($MSP_{10}$). Justification for use of halving protocols is based on a belief that the number of positive pools is unacceptably high and an expectation that if the constituents of these pools are randomly allocated to pools of half the size, a sufficient number of negative pools will be formed resulting in fewer pools that finally test positive and overall fewer tests (Munoz-Zanzi et al., 2000). It is expected that this will offset the extra cost of testing due to re-pooling before individual testing occurs. Thus a halving procedure serves as a corrective measure reducing the number of pools that are finally positive. Higher prevalences would mostly result in an increased number of positive pools and thus increased utility of a halving protocol over a non-halving simple protocol. This is likely to be the case in this situation as MH differed from MSP only by the presence of a halving procedure.

This simulation study showed somewhat paradoxically that diagnostic tests with superior sensitivity result in inferior median savings and decreased probability that savings would be at least 50% and/or 75% of the cost of individual testing (Tables 3, 4 and 7) when compared to tests with inferior sensitivity. At lower test sensitivities, a larger percentage of truly positive pools are misclassified as negative thus reducing the final number of positive pools and consequently, the ultimate number of individual tests to be performed. This savings is greater at higher prevalences and for larger pool sizes because at greater prevalences, though the percentage of misclassified results is the same, a higher absolute number of positive pools are misclassified as negative thus resulting in proportionately more savings. For a given prevalence and a pool size of 50, larger absolute numbers of samples are misclassified as negative in misclassified positive pools and hence, savings are greater than for smaller pool sizes. We attribute the better performance of larger pool sizes reported in the sensitivity analysis mainly to the decrease in sensitivity. We also note that this greater increase in performance at higher prevalences is consistent with our prior comments in this paragraph. Our results show that as expected, savings attributable to imperfect test sensitivity were achieved with the penalty of decreasing NPVs as sensitivity decreased. At prevalences of 0.01 to 0.05,

differences in NPV were small (≤ 3%) as sensitivity changed from 0.5 to 0.9 (Table 5). However, at a prevalence of 0.1 this difference was 5 % (Table 5). In a context, where investigators have a choice between diagnostic tests of greater and lesser sensitivity, depending on the consequences of false-negative results, they might consider it more cost effective to use the test of lower sensitivity if they are confident that the animal-level prevalence is ≤ 0.05.

As in all simulation studies, questions arise as to how applicable the simulated scenarios are to what is experienced under field conditions. This study was conducted for a limited number of prevalences and used the assumption of perfect test specificity for MAP. The latter assumption, which was reasonable for fecal culture for MAP, made comparisons of scenarios and interpretation of results easier. In addition, the selected values for within-herd prevalences and test sensitivity approximate current testing conditions and allow for test modifications that might increase sensitivity. However, because of the limited number of scenarios investigated we could not take into consideration all the factors that account for heterogeneity among animals in a herd or variability in laboratory handling of specimens, both of which determine the result of the testing procedures. For instance, we did not investigate scenarios in which sensitivity correlation might vary as a function of pool size. Regardless, we believe that our results provide helpful guidance for veterinarians and laboratory diagnosticians for choice of protocol and pool size when there is sound evidence that the true MAP prevalence is between 0.01 and 0.1. We also note here that our overall findings are also valid for smaller herd sizes of 100 to 200 (data not shown) albeit with much smaller differences in performances between protocols. In as much as testing conditions for other infectious diseases mimic these conditions, the results of this study would also have application to them.



## 5. Conclusions

For MAP infected herds, we advocate use of the MH protocol as a more cost-effective means of identifying infected animals than MSP and suggest that whenever there is strong evidence that there are highly infectious animals shedding large number of

MAP in feces and individual-level prevalences range from 0.01 to 0.05, pool sizes of 20 should be used. For a prevalence of 0.1, a pool size of 10 should be used and in general, for prevalences between 0.01 and 0.1, we recommend against use of pool sizes of 5. We note here that in herds that are frequently tested and in which management practices are routinely instituted, highly infected animals are likely to be culled and the prevalence of MAP likely to decrease over time (Wells et al., 2008; Ferrouillet et al., 2009). When this occurs, except when it is believed that the prevalence of MAP is very low (~ 0.01), we suggest that pool sizes of 50 be used for testing. We advocate use of MSP instead of MH only when the number of samples for testing exceeds laboratory capacity or the logistics of halving cannot be undertaken by laboratory personnel

## Acknowledgements

The study was funded by the Johne's Disease Integrated Project through USDA-CSREES-NRI grant number 2007-35204-18391.

## Conflict of Interest

None

# References


Abel, U., Schosser, R., Suss, J., 1999. Estimating the prevalence of infectious agents using pooled samples: biometrical considerations. Zentralbl. Bakteriol. 289, 550-563.

Altman, D.G., 1991. Practical Statistics for Medical Research. Chapman and Hall London.

Borel, N., Doherr, M.G., Vretou, E., Psarrou, E., Thoma, R., Pospischil, A., 2004. Seroprevalences for ovine enzootic abortion in Switzerland. Prev. Vet. Med. 65, 205-216.

Brinkhof, J.M.A., Houwers, D.J., van Maanen, C., 2007. Development of a sample pooling strategy for the serodiagnosis of small ruminant lentiviral infections using the ELITEST-MVV ELISA. Small Ruminant Res. 70, 194-199.

Christensen, J., Gardner, I.A., 2000. Herd-level interpretation of test results for epidemiologic studies of animal diseases. Prev. Vet. Med. 45, 83-106.

Collins, M.T., Gardner, I.A., Garry, F.B., Roussel, A.J., Wells, S.J., 2006. Consensus recommendations on diagnostic testing for the detection of paratuberculosis in cattle in the United States. J. Am. Vet. Med. Assoc. 229, 1912-1919.

Cowling, D.W., Gardner, I.A., Johnson, W.O., 1999. Comparison of methods for estimation of individual-level prevalence based on pooled samples. Prev. Vet. Med. 39, 211-225.

Dorfman, R., 1943. The detection of defective members of large populations. Ann. Math. Stat. 14, 436-440.

Evers, E.G., Nauta, M.J., 2001. Estimation of animal-level prevalence from pooled samples in animal production. Prev. Vet. Med. 49, 175-190.

Ferrouillet, C., Wells, S.J., Hartmann, W.L., Godden, S.M., Carrier, J., 2009. Decrease of Johne's disease prevalence and incidence in six Minnesota, USA, dairy cattle herds on a long-term management program. Prev. Vet. Med. 88, 128-137.

Gastwirth, J.L., Johnson, W.O., 1994. Screening with cost-effective quality control: Potential applications to HIV and drug testing. J. Am. Stat. Assoc. 89, 972-981.

Graff, L.E., Roeloffs, R., 1972. Group testing in the presence of test error; An extension of the Dorfman Procedure. Technom. 14, 113-122.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 45, 3-22.

Groenendaal, H., Wolf, C.A., 2008. Farm-level economic analysis of the US National Johne's Disease Demonstration Herd Project. J. Am. Vet. Med. Assoc. 233, 1852-1858.

Gupta, D., Malina, R., 1999. Group testing in presence of classification errors. Stat. Med. 18, 1049-1068.

Gwa, L.-H., Hsieh, C.-C., Ko, Y.-C., Lan, S.-J., 1992. Beyond simple pooling for HIV screening. J. Immunoassay Immunochem. 13, 545 - 557.

Jouy, E., Proux, K., Humbert, F., Rose, V., Lalande, F., Houdayer, C., Picault, J.P., Salvat, G., 2005. Evaluation of a French ELISA for the detection of *Salmonella enteritidis* and *Salmonella typhimurium* in flocks of laying and breeding hens. Prev. Vet. Med. 71, 91-103.

Kalis, C.H., Hesselink, J.W., Barkema, H.W., Collins, M.T., 2000. Culture of strategically pooled bovine fecal samples as a method to screen herds for paratuberculosis. J. Vet. Diagn. Invest. 12, 547-551.

Kennedy, J.A., Mortimer, R.G., Powers, B., 2006. Reverse transcription-polymerase chain reaction on pooled samples to detect bovine viral diarrhea virus by using fresh ear-notch-sample supernatants. J. Vet. Diagn. Invest. 18, 89-93.
Kennedy, J.A., Pearl, D., Tomky, L., Carman, J., 2008. Pooled polymerase chain reaction to detect Tritrichomonas foetus in beef bulls. J. Vet. Diagn. Invest. 20, 97-99.
Letellier, C., De Meulemeester, L., Lomba, M., Mijten, E., Kerkhofs, P., 2005. Detection of BVDV persistently infected animals in Belgium: evaluation of the strategy implemented. Prev. Vet. Med. 72, 121-125.
Maherchandani, S., Munoz-Zanzi, C.A., Patnayak, D.P., Malik, Y.S., Goyal, S.M., 2004. The effect of pooling sera on the detection of avian pneumovirus antibodies using an enzyme-linked immunosorbent assay test. J. Vet. Diagn. Invest. 16, 497-502.
Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18, 50-60.
Matsumoto, M., Nishimura, T., 1998. Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. ACM. TOMACS. 8, 3-30.
Munoz-Zanzi, C.A., Johnson, W.O., Thurmond, M.C., Hietala, S.K., 2000. Pooled-sample testing as a herd-screening tool for detection of bovine viral diarrhea virus persistently infected cattle. J. Vet. Diagn. Invest. 12, 195-203.
Munoz-Zanzi, C.A., Thurmond, M.C., Hietala, S.K., Johnson, W.O., 2006. Factors affecting sensitivity and specificity of pooled-sample testing for diagnosis of low prevalence infections. Prev. Vet. Med. 74, 309-322.
Nielsen, S.S., Toft, N., 2008. Ante mortem diagnosis of paratuberculosis: A review of accuracies of ELISA, interferon-[gamma] assay and faecal culture techniques. Vet. Microbiol. 129, 217-235.
Patel, M.S., 1962. Group-screening with more than two stages. Technom. 4, 209-217.
R, Development Core Team 2005. R: A language and environment for statistical computing. Vienna, Austria.
Rovira, A., Clement, T., Christopher-Hennings, J., Thompson, B., Engle, M., Reicks, D., Munoz-Zanzi, C., 2007. Evaluation of the sensitivity of reverse-transcription polymerase chain reaction to detect porcine reproductive and respiratory syndrome virus on individual and pooled samples from boars. J. Vet. Diagn. Invest. 19, 502-509.
Sanderson, M.W., Sargeant, J.M., Nagaraja, T.G., 2005. Effect of pooling bovine fecal samples on the sensitivity of detection of E-coli O157 : H7. Vet. Microbiol. 110, 125-130.
Singer, R.S., Cooke, C.L., Maddox, C.W., Isaacson, R.E., Wallace, R.L., 2006. Use of pooled samples for the detection of Salmonella in feces by polymerase chain reaction. J. Vet. Diagn. Invest. 18, 319-325.
Sterrett, A., 1957. On the detection of defective members of large populations. Ann. Math. Stat. 28, 1033-1036.
Toribio, J.A., Sergeant, E., 2007. A comparison of methods to estimate the prevalence of ovine Johne's infection from pooled faecal samples. Aust. Vet. J. 85, 317-324.
van Schaik, G., Stehman, S.M., Schukken, Y.H., Rossiter, C.R., Shin, S.J., 2003. Pooled fecal culture sampling for Mycobacterium avium subsp. paratuberculosis at different herd sizes and prevalence. J. Vet. Diagn. Invest. 15, 233-241.
van Schaik, G., Pradenas F, M., Mella N, A., Kruze V, J., 2007. Diagnostic validity and costs of pooled fecal samples and individual blood or fecal samples to determine the

cow- and herd-status for Mycobacterium avium subsp. paratuberculosis. Prev. Vet. Med. 82, 159-165.
Wells, S.J., Hartmann, W.L., Anderson, P.L., 2008. Evaluation of progress made by dairy and beef herds enrolled in the Minnesota Johne's Disease Control Program. J. Am. Vet. Med. Assoc. 233, 1920-1926.
Williams, C.J., Moffitt, C.M., 2005. Estimation of pathogen prevalence in pooled samples using maximum likelihood methods and open-source software. J. Aquat. Anim. Health 17, 386-391.