Discrete Compound Tests and Dorfman's Methodology in the Presence of Misclassification

Rui Santos, João Paulo Martins, and Miguel Felgueiras

Abstract Compound tests can be used to save resources for classification or estimation purposes in clinical trials and quality control. Nevertheless, the usually applied methodologies are restricted to qualitative group tests. Moreover, when quantitative compound tests are applied the problem is to ascertain whether the amount of some substance of any individual in the group is greater or lower than a prefixed threshold. An overview of discrete compound tests application highlights the advantages (to save resources) and disadvantages (higher probability of misclassification), and suggests criteria to assess the suitability of applying Dorfman's methodology.

1 Introduction

Let p be the prevalence rate of some infection in a population with N individuals and the Bernoulli trials X_i, $i = 1, \ldots, N$, denote the presence $(X_i = 1)$ and absence $(X_i = 0)$ of the infection in the i-th member of the population, i.e. $X_i \sim \text{Ber}(p)$. Moreover, the random variables (r.v.s) X_i will be consider mutually independent due to the random sampling. Thus, the number of infected members in a group of size n is a r.v. with binomial distribution, i.e. $\text{I}^{[n]} \sim \text{Bin}(n, p)$. In addition, suppose that the identification of the infection is carried out by a clinical trial by counting the number of a certain type of bacteria in a milliliter of blood. If this number is greater than a given threshold t the individual is classified as infected, otherwise is classified as not infected (the opposite inequalities could also be applied with the appropriate adaptations). Furthermore, let the number of bacteria in one *ml* of blood be given

Rui Santos and João Paulo Martins
School of Technology and Management, Polytechnic Institute of Leiria, CEAUL — Center of Statistics and its Applications, e-mail: rui.santos@ipleiria.pt, jpmartins@ipleiria.pt

Miguel Felgueiras
School of Technology and Management, Polytechnic Institute of Leiria, CEAUL — Center of Statistics and its Applications, CIIC — Computer Science and Communications Research Centre of Polytechnic Institute of Leiria, e-mail: mfelg@ipleiria.pt

by the r.v. Y_i. If $X_i = 0$ (uninfected individual), then $Y_i = Y_i^- \sim \mathbf{D}_0(\boldsymbol{\theta_0})$ where $\mathbf{D}_0$ denotes some count distribution with support $\mathrm{S}_0 \in \mathbb{N}_0$ and parameters vector $\boldsymbol{\theta_0}$. If $X_i = 1$ (infected individual), then $Y_i = Y_i^+ \sim \mathbf{D}_1(\boldsymbol{\theta_1})$ where $\mathbf{D}_1$ is some count distribution with support $\mathrm{S}_1 \in \mathbb{N}_0$ and parameters vector $\boldsymbol{\theta_1}$. In fact, the r.v.s Y^- and Y^+ can have the same support (at least partly, i.e. $\mathrm{S}_0 \cap \mathrm{S}_1 = \mathrm{S} \neq \emptyset$). Furthermore, any classification methodology has a nonzero probability to perform an erroneous classification for those values in S, which leads to the presence of misclassification in the individual test (if $\mathrm{S} = \emptyset$ then there would be no problem of misclassification). Nevertheless, Y_i^- and Y_i^+ must have some relationship to ensure that the probability of misclassification is small, for instance Y_i^+ shall stochastically dominates Y_i^- (i.e. $F_{Y^+}(x) \leq F_{Y^-}(x), \forall x \in \mathbb{R}$), otherwise the test would not have much information concerned with the infection. In some applications $\mathbf{D}_0$ and $\mathbf{D}_1$ denote the same distribution with a change of location and a possible alteration of scale. The usual applied distributions are the Poisson, the negative binomial, and the binomial distributions. The previously illustrated context will be used throughout this work in the description of the proposed methodologies. However, the applicability of these methods is not restricted to blood analysis, such as the screening of infectious diseases like HIV (see [25, 26]), since it can easily be adjusted to be applied to any fluid that can be mixed, e.g. in industrial quality control cf. [3, 10].

An overview of discrete compound tests application will be given in this section. It is described how to perform compound tests, including its goals (estimation and classification) as well as the main usual measures applied to evaluate the accuracy of the test results. Hereafter, a new evaluation measure of the application of compound tests in the presence of misclassification is proposed in order to identify the situations in which the compound test can be applied without an excessive probability of misclassification (section 2). Section 3 provides a detailed description of two different methodologies to perform the compound tests and, therefore, to set up its cut off points. In section 4 the performance of the two proposed methodologies is investigated via simulation. Finally, the main conclusions are outlined in section 5.

1.1 Group/Compound Tests

In an individual test we are performing the statistical hypothesis test formalized by $\mathbf{H_0} : X_i = 0$ versus $\mathbf{H_1} : X_i = 1$. If the cut off point is the threshold t then the significant level is $\alpha = \mathrm{P}(Y_i > t | X_i = 0) = \mathrm{P}\left(Y_i^- > t\right)$ and the power of the test $1 - \beta = \mathrm{P}(Y_i > t | X_i = 1) = \mathrm{P}\left(Y_i^+ > t\right)$. To perform a compound test, one *ml* of blood is taken from each of the n members of the group and then mixed. Thus, we get n *ml* of pooled blood where the number of bacteria is given by $B_n = \sum_{i=1}^{n} Y_i$. After being turned into a totally homogeneous fluid, it is withdraw, randomly, one *ml* of this pooled fluid to test. The number B_1 of bacteria in this *ml* of fluid can be computed using hierarchical models, more precisely by the application of a binomial filter taking into account that each of the B_n bacterias in the n *ml* of blood has probability n^{-1} of being in the chosen *ml*. Therefore, $B_1 \sim \mathrm{Bin}\left(B_n, \frac{1}{n}\right)$, cf. [21]. This *ml* of pooled blood will be used to perform the compound test. The main idea

of this test is to identify if there is any infected member in the group. Thus, if the test results negative then all member in the group are uninfected. Otherwise, at least one member of the group must be infected. Therefore, the statistical hypothesis to perform are $\mathbf{H_0} : \sum_{i=1}^{n} X_i = 0$ versus $\mathbf{H_1} : \sum_{i=1}^{n} X_i \geq 1$. In fact, it aims to identify if any of the individuals in the group $(i = 1, \ldots, n)$ can be classified as infected, i.e. if $\max(Y_1, \ldots, Y_n) > t$, using the only available information B_1 – the number of bacteria in the *ml* of pooled blood. The significant level of the compound test is $\alpha = \mathrm{P}(B_i > t^* | \sum_{i=1}^{n} X_i = 0)$ and the power of the test $1 - \beta = \mathrm{P}(B_i > t^* | \sum_{i=1}^{n} X_i \geq 1)$, where t^* denotes the applied cut off point of the compound test.

1.2 Classification and Prevalence Rate Estimation

The use of compound test has two main goals: classification (categorization of each individual as infected or not infected) and the estimation of the prevalence rate p. Dorfman's methodology was first used in the second World War in order to identify all American soldiers infected with syphilis, cf. [5]. In this classification methodology all population is divided in groups with n individuals and a compound test is performed to each group. As outlined in the previous subsection, if the result is negative then all members are classified as not infected, otherwise at least one member should be classified as infected. Consequently, individual tests are performed in order to identify which elements are indeed infected. The main goal is to compute the optimal dimension n (according to the prevalence rate p) in order to minimize the expected number of tests required to identify all infected individuals. Whereas in individual tests it is required to perform N tests to classify all N individuals in the population, in Dorfman's methodology the expected number of tests is $\mathrm{E}[T_n] = N\left(\frac{1}{n} + 1 - (1-p)^n\right)$. Hence, the relative cost (expected number of test for the classification of each individual) is $\mathrm{RC} = \frac{n+1}{n} - (1-p)^n$, $n \geq 2$, which can be used to establish the optimal size n for each prevalence rate p. If $p \leq 0.3$ then $\mathrm{RC} \leq 1$ and, therefore, is better to use the Dorfman's methodology than individual tests. The optimal size n for some prevalence rates p can be found in [5, 21] and elsewhere. For $p \leq 0.12$ it can be used a linear approximation and the optimal group size can be approximate by $n \approx \frac{1}{\sqrt{p}} + 0.5$ with good results, cf. [6]. In this work it is considered that the cost of mixing samples is negligible [16] and therefore the number of tests matches to the primary cost related to the classification procedure.

The first few classification methodologies considered the absence of misclassification and were restricted to qualitative group tests (identification of the presence or absence of some substance in the compound fluid). Subsequently, new algorithms have been proposed in order to minimize the number of tests required for the correct classification of all individuals in the population, mainly applying halving nested procedures or hierarchical algorithms (generalizations of the Dorfman's methodology in which positive groups are repeatedly divided into smaller non-overlapping subgroups until all members be individually tested, cf. [6, 8, 13, 15, 23, 24]), square array testing (with the use of overlapping pools, cf. [12, 19, 28]), and multidimensional array algorithms (an extension to higher dimensional arrays, cf. [1, 20]).

In addition, the use of compound tests can also be useful in the estimation of the prevalence rate p as [22] exhibit. Under certain conditions, these estimators have better performance than the estimators based on individual tests, allowing to reduce the number of performed tests and simultaneously to achieve more accurate estimates with respect to the bias, efficiency as well as robustness, cf. [4, 7, 9, 14, 17, 22] among others. Some packages with applications of several compound testing estimators, such as *binGroup* for the R software [2], are available.

Nevertheless, in both cases (classification and estimation) the use of compound tests should only be performed for low prevalence rates, as otherwise the individual tests outperforms. In this work the compound tests will be used for classification purposes applying the Dorfman's methodology.

1.3 Misclassification Evaluation

The usual applied measure to evaluate the misclassification problem are the specificity, the sensitivity, the positive predictive value, and the negative predictive value. These measures can be defined for the individual test, for the compound test, and for the application of a specific methodology for classification.

1.3.1 Misclassification in Individual Tests

Performing individual tests, the individual specificity is the probability of getting a negative result (X_i^-) from a not infected individual, i.e. $\varphi_e = \mathrm{P}\left(X_i^- \middle| X_i = 0\right)$, and the individual sensitivity is the probability of getting a positive result (X_i^+) from an infected individual, i.e. $\varphi_s = \mathrm{P}\left(X_i^+ \middle| X_i = 1\right)$. The positive predictive value is the probability of having an infected sample in a positive individual test, i.e. $\mathrm{PPV} = \mathrm{P}\left(X_i = 1 \middle| X_i^+\right)$, and the negative predictive value is the probability of having an uninfected sample in a negative individual test, i.e. $\mathrm{NPV} = \mathrm{P}\left(X_i = 0 \middle| X_i^-\right)$.

1.3.2 Misclassification in Compound Tests

Let us now consider compound tests performed to groups with size n, and let $X^{[+,n]}$ and $X^{[-,n]}$ denote respectively a positive and a negative compound result. Hence, the compound specificity is given by $\varphi_e^{[n]} = \mathrm{P}(X^{[-,n]} \middle| \mathrm{I}^{[n]} = 0)$, and the compound sensitivity by $\varphi_s^{[n]} = \mathrm{P}(X^{[+,n]} \middle| \mathrm{I}^{[n]} \geq 1)$. Nevertheless, $\varphi_s^{[n]}$ depends of the number of infected members in the group due to the dilution and consequent rarefaction of the number of bacteria. Thus, setting $\varphi_s^{[j,n]} = \mathrm{P}(X^{[+,n]} \middle| \mathrm{I}^{[n]} = j)$, the rarefaction factor can be added in the $\varphi_s^{[n]}$ computation, by doing $\varphi_s^{[n]} = \sum_{j=1}^{n} \varphi_s^{[j,n]} \mathrm{P}(\mathrm{I}^{[n]} = j \middle| \mathrm{I}^{[n]} \geq 1)$, cf. [21]. Moreover, $\varphi_s^{[n]} \approx \varphi_s^{[1,n]}$ for low prevalence rates (see [21]). Similarly, the compound positive predictive value is $\mathrm{PPV}^{[n]} = \mathrm{P}(\sum_{i=1}^{n} X_i \geq 1 \middle| X^{[+,n]})$ and the compound negative predictive value $\mathrm{NPV}^{[n]} = \mathrm{P}(\sum_{i=1}^{n} X_i = 0 \middle| X^{[-,n]})$. Generally, the compound sensitivity decreases as the group size increases, due to the dilution. In the literature there are different procedures in order to model the dilution factor,

cf. [11, 21, 27, 29]. The selection of the most suitable procedure for group testing depends on the sensitivity, the specificity and the monetary costs of the process [16].

1.3.3 Misclassification in Classification Methodologies

The misclassification measures previously defined can be generalized in order to measure the misclassification in some classification methodology $\mathscr{M}$. Thus, the same definitions are applied as in the individual tests, but the probabilities are computed taking into consideration the application of the methodology under investigation. Hence, the $\mathscr{M}$ methodology specificity is the probability of an uninfected individual be classified as uninfected by the application of the $\mathscr{M}$ methodology. Thus, the Dorfman's methodology specificity is given by (cf. [21])

$$\varphi_{e_n} = \mathrm{P}\left(X_i^- \mid X_i = 0\right) = \sum_{i=0}^{n-1} \mathrm{P}\left(X_1^- \mid X_1 = 0, \mathrm{I}^{[n-1]} = i\right) \mathrm{P}\left(\mathrm{I}^{[n-1]} = i\right), \qquad (1)$$

and, analogously, the Dorfman's methodology sensitivity is given by

$$\begin{aligned} \varphi_{s_n} &= \mathrm{P}\left(X_1^+ \mid X_1 = 1\right) = \sum_{i=0}^{n-1} \mathrm{P}\left(X_1^+ \mid X_1 = 1, \mathrm{I}^{[n-1]} = i\right) \mathrm{P}\left(\mathrm{I}^{[n-1]} = i\right) \\ &= \varphi_s \sum_{i=0}^{n-1} \varphi_s^{[i+1,n]} \mathrm{P}\left(\mathrm{I}^{[n-1]} = i\right). \end{aligned} \qquad (2)$$

Furthermore, similar reasoning can be applied to the Dorfman's positive predictive value $\mathrm{PPV}_{\mathscr{D}_n}$ and Dorfman's negative predictive value $\mathrm{NPV}_{\mathscr{D}_n}$.

2 A Proposal to Measure the Quality of the Individual Test

In carrying out individual tests, if the threshold t increases then the specificity increases and the sensitivity decreases. Hence, if we improve the specificity (sensitivity) then the sensitivity (specificity) gets worse analogously to the relation between α (probability of error type I) and β (probability of error type II) in a statistical hypothesis test. Hence, the threshold t allows the adjustment of the balance between sensitivity and specificity and, thus, a threshold t^e can be set in order to equalize the sensitivity to the specificity (or, if not possible, to minimize its distance). The value of these probabilities, denoted by ϕ, defines a measure of the quality of the individual test performance.

Definition 1. In the individual test, the probability ϕ which verifies $\varphi_s = \varphi_e = \phi$ for some threshold t^e is the quality measure of the individual test performance QMITP.

A high value of ϕ (in the neighborhood of the unit, $\phi \approx 1$) implies that the hypothesis test has a low probability of misclassification. Otherwise, if ϕ is much lower than 1, it implies a high probability of misclassification.

In fact, some information is lost when bloods are mixed. Hence, the compound tests should not be applied in cases in which ϕ is quite low, because the mixture will

still further increase the probability of misclassification. Subsequently, compound tests can only be applied if the individual tests have a good performance. Therefore, in section 4 this quality measure of the individual test performance will be used in simulation in order to assess if it can be also used to measure the adequacy of the application of compound tests.

3 Methodologies for the compound tests

In this section two different methodologies to perform the compound tests are described. As in statistical hypothesis tests, it is impossible to improve the two probabilities of misclassification and therefore only one can actually be controlled. Thus, each of the applied methodology pretends to control one of these probabilities. Moreover, the first methodology (the usual applied methodology which will be denoted by $\mathbf{M_1}$) controls the compound specificity insofar the second methodology ($\mathbf{M_2}$) establishes the compound sensitivity.

3.1 The Usual Methodology – $\mathbf{M_1}$

The usual applied hypothesis test (first methodology – $\mathbf{M_1}$) is, cf. [21],

$$\mathbf{H_0}: \sum_{i=1}^{n} X_i = 0 \ \left(\mathrm{I}^{[n]} = 0\right) \ \text{versus} \ \mathbf{H_1}: \sum_{i=1}^{n} X_i \geq 1 \ \left(\mathrm{I}^{[n]} \geq 1\right).$$

Hence, the test size is given by $\alpha = \mathrm{P}(X^{[+,n]} | \sum_{i=1}^{n} X_i = 0) = 1 - \varphi_e^{[n]}$ and, therefore, the compound specificity is fixed. Thus, the specificity is controlled by setting the value of α, but it neglects the sensitivity, i.e. the occurrence of false negatives. Moreover, by (1) the φ_{e_n} is equal to (see [21])

$$\mathrm{P}\left(\mathrm{I}^{[n-1]} = 0\right)\left[\varphi_e^{[n]} + (1-\varphi_e^{[n]})\varphi_e\right] + \sum_{i=1}^{n-1} \mathrm{P}\left(\mathrm{I}^{[n-1]} = i\right)\left[\varphi_s^{[i,n]}\varphi_e + \left(1-\varphi_s^{[i,n]}\right)\right]$$

and, therefore, $\varphi_{e_n} \geq \varphi_e^{[n]}$. Hence, the Dorfman's methodology specificity verifies $\varphi_{e_n} \geq 1-\alpha$ and the size of test sets a lower limit for φ_{e_n}.

3.2 An Alternative Methodology – $\mathbf{M_2}$

In most applications, it is essential to control the occurrence of false negative results (i.e., to set the sensitivity). With this goal we propose an alternative methodology $\mathbf{M_2}$, which corresponds to the hypothesis test

$$\mathbf{H_0}: \sum_{i=1}^{n} X_i \geq 1 \ \left(\mathrm{I}^{[n]} \geq 1\right) \ \text{versus} \ \mathbf{H_1}: \sum_{i=1}^{n} X_i = 0 \ \left(\mathrm{I}^{[n]} = 0\right).$$

The test size α is given by $\alpha = \mathrm{P}(X^{[-,n]} | \sum_{i=1}^{n} X_i \geq 1) = 1 - \varphi_s^{[n]}$. Hence, the compound sensitivity is fixed and, therefore, the probability of false negative results is controlled. A seeming drawback of $\mathbf{M_2}$ comparing with $\mathbf{M_1}$ is the complexity of the cut off point computation due to the different scenarios in $\mathbf{H_0}$. However, in practice, to compute the cut off point it can be considered the following hypothesis test

$$\mathbf{H_0} : \sum_{i=1}^{n} X_i = 1 \quad \left(\mathrm{I}^{[n]} = 1\right) \quad \text{versus} \quad \mathbf{H_1} : \sum_{i=1}^{n} X_i = 0 \quad \left(\mathrm{I}^{[n]} = 0\right).$$

The results of applying this simplified $\mathbf{M_2}$ are quite similar to $\mathbf{M_2}$ because the probability of getting more than one infected individual in the group is quite low (as previously stated for low prevalence rates, as is to be considered for the use of compound tests). On the other hand, in this simplified $\mathbf{M_2}$ the significance level is given by $\alpha = \mathrm{P}(X^{[-,n]} | \sum_{i=1}^{n} X_i = 1) = 1 - \varphi_s^{[1,n]}$, and therefore α will set $\varphi_s^{[1,n]}$. In addition, as having just one infected individual in the group corresponds to the worst case scenario, then $\varphi_s^{[1,n]} \leq \varphi_s^{[2,n]} \leq \cdots \leq \varphi_s^{[n,n]}$, and consequently $\varphi_s^{[1,n]} \leq \varphi_s^{[n]}$. Thus, the α value will set up a lower limit for the compound sensitivity $\varphi_s^{[n]}$. Nevertheless, by (2) the Dorfman's sensitivity is lower than the compound sensitivity $\varphi_{s_n} \leq \varphi_s^{[n]}$. Hence, the Dorfman's sensitivity can be lower or higher than $1 - \alpha$. This simplified methodology was already proposed in [18] but without any examination, which will be carried out in the simulations performed in the section 4.

4 Simulation

The main goal of this simulations is to investigate the use of the second simplified methodology as well as the quality measure of the individual test performance ϕ.

4.1 Simulation Settings

All simulations were performed in software R using 10^6 groups in each simulation and applying different prevalence rates p, significance levels α, group dimensions n and QMITP ϕ. The case $n = 1$ corresponds to the restricted use of individual tests. For an infected individual the distribution $\mathbf{D}_1$ of Y_i^+ was defined through a change of location $Y_i^+ = \mu' + Y_i^-$ in which μ' is computed in order to verify a specific value for ϕ. The investigated measures were the Dorfman's sensitivity φ_{s_n}, specificity φ_{e_n}, positive $\mathrm{PPV}_{\mathscr{D}_n}$ and negative $\mathrm{NPV}_{\mathscr{D}_n}$ predictive values, and the relative cost RC.

4.2 Results and Discussion

Table 1 shows the results of the application of both hypothesis test methodologies, using a significance level of 5%, a prevalence rate of 1% and $Y_i^- \sim \mathrm{Poisson}(100)$. In addition, it were analyzed multiple groups dimensions although for $p = 0.01$

Table 1 Simulations with $\alpha = \mathbf{0.05}$, $p = 0.01$ $(n^* = 11)$, $Y_i^- \sim \text{Poisson}(100)$, and 10^6 groups

	Methodology $\mathbf{M_1}$					Methodology $\mathbf{M_2}$				
	φ_{s_n}	φ_{e_n}	$\text{PPV}_{\mathscr{D}_n}$	$\text{NPV}_{\mathscr{D}_n}$	RC	φ_{s_n}	φ_{e_n}	$\text{PPV}_{\mathscr{D}_n}$	$\text{NPV}_{\mathscr{D}_n}$	RC
					$\phi = \mathbf{0.95}$					
$n = 1$	94.11	**95.71**	18.02	99.94	100	**94.11**	95.71	18.02	99.94	100
$n = 2$	71.93	**98.29**	29.65	99.71	55.73	**91.33**	96.33	20.00	99.91	**74.10**
$n = 3$	57.65	**98.66**	30.25	99.57	39.57	**91.29**	96.19	19.51	99.91	74.42
$n = 5$	44.39	**98.90**	29.02	99.44	27.15	**90.44**	96.16	19.20	99.90	77.86
$n = 7$	35.05	**99.10**	28.36	99.34	21.05	**90.12**	96.13	19.07	99.90	80.20
$n = 10$	28.51	**99.23**	27.15	99.28	16.77	**90.29**	96.07	18.81	99.90	84.16
$n = 20$	22.44	**99.30**	24.58	99.22	12.99	**90.37**	96.01	18.63	99.90	88.66
$n = 30$	18.36	**99.40**	23.56	99.18	**11.02**	**90.70**	95.97	18.48	99.90	90.94
$n = 50$	17.89	**99.37**	22.22	99.17	11.31	**91.16**	95.91	18.36	99.91	93.40
$n = 100$	16.81	**99.36**	21.06	99.16	11.78	**91.48**	95.88	18.32	99.91	95.12
					$\phi = \mathbf{0.99}$					
$n = 1$	99.88	**95.71**	19.26	100	100	**94.26**	99.88	88.99	99.94	100
$n = 2$	95.33	**98.13**	34.15	99.95	56.82	**94.66**	98.29	35.94	99.94	56.17
$n = 3$	84.78	**98.67**	39.22	99.84	40.30	**94.89**	97.58	28.45	99.95	**48.79**
$n = 5$	69.06	**98.93**	38.89	99.66	27.85	**95.01**	96.92	23.83	99.95	54.86
$n = 7$	56.15	**99.06**	37.67	99.55	22.45	**95.65**	96.61	22.20	99.95	63.57
$n = 10$	48.45	**99.09**	34.89	99.48	19.22	**95.18**	96.48	21.41	99.95	69.21
$n = 20$	34.16	**99.23**	31.01	99.33	14.55	**95.96**	96.19	20.27	99.96	82.07
$n = 30$	30.16	**99.25**	28.79	99.30	13.77	**96.19**	96.09	19.86	99.96	86.46
$n = 50$	26.65	**99.26**	26.69	99.26	**13.53**	**96.44**	96.01	19.63	99.96	90.23
$n = 100$	25.69	**99.19**	24.22	99.25	15.37	**97.32**	95.90	19.35	99.97	94.45
					$\phi = \mathbf{0.999}$					
$n = 1$	100	**95.70**	19.37	100	100	**94.00**	100	99.83	99.94	100
$n = 2$	99.61	**98.28**	36.98	100	56.26	**94.87**	99.68	75.00	99.95	52.29
$n = 3$	97.17	**98.57**	40.71	99.97	41.21	**94.48**	99.02	49.35	99.94	38.88
$n = 5$	86.32	**98.92**	44.81	99.86	28.83	**94.82**	98.07	33.33	99.95	**36.82**
$n = 7$	75.01	**99.03**	43.98	99.75	23.75	**95.00**	97.53	28.00	99.95	43.23
$n = 10$	62.65	**99.11**	41.62	99.62	20.00	**95.14**	97.11	24.96	99.95	52.43
$n = 20$	46.52	**99.14**	35.56	99.46	16.69	**95.94**	96.50	21.75	99.96	72.12
$n = 30$	39.71	**99.17**	32.53	99.39	**15.66**	**96.33**	96.28	20.68	99.96	80.31
$n = 50$	36.82	**99.09**	29.06	99.36	17.02	**96.57**	96.12	20.09	99.96	86.68
$n = 100$	35.09	**98.98**	25.79	99.34	19.80	**97.92**	95.91	19.49	99.98	94.05

the more efficient size in Dorfman's methodology (without misclassification) is 11 individuals in each group ($n^* = 11$). Nevertheless, the efficient size can correspond to a case with high probability of misclassification which shall be avoid.

The simulation results clearly show that the two methodologies fulfill their goals: whereas the significance level in $\mathbf{M_1}$ controls φ_{e_n}, implying $\varphi_{e_n} \geq 1 - \alpha$ and in most cases $\varphi_{e_n} \approx 0.99$, in $\mathbf{M_2}$ controls φ_{s_n}, despite of φ_{s_n} having a higher variability in the case of $\phi = 0.95$. Besides, the observed φ_{s_n} converge quickly to zero whenever n increases in $\mathbf{M_1}$, but φ_{e_n} still achieve good results in $\mathbf{M_2}$ even when n increases. The $\text{NPV}_{\mathscr{D}_n}$ values are quite reasonable, but $\text{PPV}_{\mathscr{D}_n}$ do not get so good performance. These results are a consequence of working with low prevalence rates, and as such,

Table 2 Simulations with $\phi = \mathbf{0.99}$ and $Y_i^- \sim \text{Poisson}(100)$

	Methodology $\mathbf{M_1}$					Methodology $\mathbf{M_2}$				
	φ_{s_n}	φ_{e_n}	$\text{PPV}_{\mathscr{D}_n}$	$\text{NPV}_{\mathscr{D}_n}$	RC	φ_{s_n}	φ_{e_n}	$\text{PPV}_{\mathscr{D}_n}$	$\text{NPV}_{\mathscr{D}_n}$	RC
				$\alpha = \mathbf{0.10}$, and $p = 0.01$						
$n = 3$	92.38	**96.40**	20.61	99.92	45.58	**89.85**	97.06	23.59	99.89	42.96
$n = 5$	79.19	**97.10**	21.69	99.78	32.93	**89.68**	95.50	16.82	99.89	42.61
$n = 10$	59.02	**97.62**	20.01	99.58	23.66	**90.38**	93.94	13.07	99.90	55.28
$n = 20$	46.83	**97.72**	17.17	99.45	20.76	**91.07**	93.04	11.66	99.90	68.31
				$\alpha = \mathbf{0.05}$, and $p = 0.01$						
$n = 3$	84.78	**98.67**	39.22	99.84	40.30	**94.89**	97.58	28.45	99.95	48.79
$n = 5$	69.06	**98.93**	38.89	99.66	27.85	**95.01**	96.92	23.83	99.95	54.86
$n = 10$	48.45	**99.09**	34.89	99.48	19.22	**95.18**	96.48	21.41	99.95	69.21
$n = 20$	34.16	**99.23**	31.01	99.33	14.55	**95.96**	96.19	20.27	99.96	82.07
				$\alpha = \mathbf{0.01}$, and $p = 0.01$						
$n = 3$	62.76	**99.83**	78.50	99.62	36.10	**98.14**	99.19	55.26	99.98	66.37
$n = 5$	40.97	**99.88**	77.23	99.41	22.82	**98.05**	99.17	54.38	99.98	76.30
$n = 10$	23.16	**99.91**	72.96	99.23	12.81	**98.21**	99.15	53.88	99.98	91.66
$n = 20$	16.84	**99.92**	68.24	99.17	08.48	**98.28**	99.14	53.57	99.98	96.78
				$p = \mathbf{0.05}$ and $\alpha = 0.05$						
$n = 3$	85.83	**98.44**	74.45	99.25	49.44	**95.06**	97.44	66.22	99.73	58.08
$n = 5$	73.82	**98.43**	71.18	98.62	40.11	**96.26**	96.68	60.33	99.80	67.42
$n = 10$	60.22	**98.42**	66.60	97.92	33.71	**96.48**	96.28	57.60	99.81	80.82
$n = 20$	57.95	**98.20**	62.91	97.79	36.11	**98.01**	95.95	56.00	99.89	93.50
				$p = \mathbf{0.01}$ and $\alpha = 0.05$						
$n = 3$	84.78	**98.67**	39.22	99.84	40.30	**94.89**	97.58	28.45	99.95	48.79
$n = 5$	69.06	**98.93**	38.89	99.66	27.85	**95.01**	96.92	23.83	99.95	54.86
$n = 10$	48.45	**99.09**	34.89	99.48	19.22	**95.18**	96.48	21.41	99.95	69.21
$n = 20$	34.16	**99.23**	31.01	99.33	14.55	**95.96**	96.19	20.27	99.96	82.07
				$p = \mathbf{0.001}$, and $\alpha = 0.05$						
$n = 3$	83.49	**98.72**	6.19	99.98	38.16	**94.28**	97.60	3.82	99.99	46.63
$n = 5$	66.27	**99.04**	6.38	99.97	25.10	**94.79**	96.95	2.97	99.99	52.07
$n = 10$	44.04	**99.31**	6.02	99.94	15.30	**94.59**	96.54	2.67	99.99	65.76
$n = 20$	27.18	**99.48**	4.93	99.93	10.34	**95.00**	96.28	2.49	99.99	77.77

there are many uninfected groups and few infected ones in the 10^6 simulated. In $\mathbf{M_1}$ it can be achieved a very low RC but with high probability of misclassification, while in $\mathbf{M_2}$ the efficiency is not so good but both φ_{s_n} and φ_{s_n} have good performance.

Table 2 investigates different prevalence rates $p \in \{0.05, 0.01, 0.001\}$ and multiple significance levels $\alpha \in \{0.1, 0.05, 0.01\}$ with $\phi = 0.99$. The results are as expected, i.e. the φ_{e_n} (φ_{s_n}) decreases and the φ_{s_n} (φ_{e_n}) increases in methodology $\mathbf{M_1}$ ($\mathbf{M_2}$) when the significance level increases. Moreover, the use of different prevalence rates (all used rates are low, because compound test should not be used otherwise) does not seem to have a major impact on the sensitivity and specificity.

Different distributions (Poisson, negative binomial, and binomial) and different parameters values (maintaining the same expected value in the distinct distributions)

Table 3 Simulations with $p = 0.01$, $\phi = \mathbf{0.99}$, $\alpha = \mathbf{0.05}$

	Methodology $\mathbf{M_1}$					Methodology $\mathbf{M_2}$				
	φ_{s_n}	φ_{e_n}	$\text{PPV}_{\mathscr{D}_n}$	$\text{NPV}_{\mathscr{D}_n}$	RC	φ_{s_n}	φ_{e_n}	$\text{PPV}_{\mathscr{D}_n}$	$\text{NPV}_{\mathscr{D}_n}$	RC
					$Y_i^- \sim \text{Poisson}(10)$					
$n = 3$	84.52	**98.49**	36.11	99.84	40.33	**94.70**	97.34	26.42	99.95	47.77
$n = 5$	64.21	**98.87**	36.42	99.64	27.17	**93.80**	96.77	22.67	99.94	49.12
$n = 10$	43.99	**99.08**	32.70	99.43	17.88	**94.70**	96.08	19.65	99.94	65.49
					$Y_i^- \sim \text{Negative Binomial}(1, \frac{1}{11})$					
$n = 3$	75.90	**98.19**	29.84	99.75	40.45	**94.48**	97.25	25.88	99.94	45.11
$n = 5$	53.83	**98.56**	27.38	99.53	27.40	**94.89**	96.52	21.52	99.95	45.15
$n = 10$	37.03	**98.90**	25.36	99.36	17.74	**95.34**	96.00	19.41	99.95	59.94
					$Y_i^- \sim \text{Binomial}(20, \frac{1}{2})$					
$n = 3$	82.34	**99.23**	51.83	99.82	40.27	**92.44**	98.73	42.33	99.92	47.95
$n = 5$	63.05	**99.45**	53.58	99.62	27.26	**93.20**	98.46	37.99	99.93	54.00
$n = 10$	40.25	**99.60**	50.55	99.40	17.14	**94.79**	98.21	34.83	99.95	71.92
					$Y_i^- \sim \text{Poisson}(100)$					
$n = 3$	84.78	**98.67**	39.22	99.84	40.30	**94.89**	97.58	28.45	99.95	48.79
$n = 5$	69.06	**98.93**	38.89	99.66	27.85	**95.01**	96.92	23.83	99.95	54.86
$n = 10$	48.45	**99.09**	34.89	99.48	19.22	**95.18**	96.48	21.41	99.95	69.21
					$Y_i^- \sim \text{Negative Binomial}(1, \frac{1}{101})$					
$n = 3$	74.54	**98.20**	29.34	99.74	40.21	**94.66**	97.16	25.05	99.94	45.21
$n = 5$	53.63	**98.54**	27.15	99.53	27.37	**94.96**	96.43	21.23	99.95	44.99
$n = 10$	37.34	**98.85**	24.63	99.37	17.89	**95.07**	95.84	18.72	99.95	59.51
					$Y_i^- \sim \text{Binomial}(200, \frac{1}{2})$					
$n = 3$	82.85	**98.79**	40.87	99.82	40.33	**94.44**	97.66	28.98	99.94	51.03
$n = 5$	64.83	**99.05**	40.85	99.64	27.61	**94.68**	97.15	25.12	99.94	57.03
$n = 10$	46.03	**99.20**	36.74	99.46	18.83	**95.13**	96.78	22.95	99.95	71.48
					$Y_i^- \sim \text{Poisson}(1000)$					
$n = 3$	84.84	**98.51**	36.65	99.84	40.48	**94.93**	97.15	25.27	99.95	50.40
$n = 5$	68.91	**98.75**	35.59	99.69	28.36	**95.15**	96.45	21.19	99.95	56.28
$n = 10$	46.51	**99.04**	32.92	99.46	18.61	**95.01**	95.97	19.26	99.95	69.48
					$Y_i^- \sim \text{Negative Binomial}(1, \frac{1}{1001})$					
$n = 3$	77.40	**98.12**	29.58	99.77	40.54	**95.03**	97.14	25.26	99.95	45.23
$n = 5$	53.26	**98.56**	27.37	99.52	27.22	**95.46**	96.36	21.01	99.95	45.65
$n = 10$	37.13	**98.85**	24.61	99.36	17.79	**95.68**	95.78	18.61	99.95	60.45
					$Y_i^- \sim \text{Binomial}(2000, \frac{1}{2})$					
$n = 3$	84.87	**98.56**	37.31	99.85	40.52	**94.66**	97.31	26.24	99.94	49.95
$n = 5$	68.09	**98.85**	37.50	99.67	28.04	**94.53**	96.71	22.56	99.94	54.33
$n = 10$	48.80	**99.02**	33.56	99.48	19.39	**94.91**	96.19	20.15	99.95	69.12

were analyzed in Table 3 in order to evaluate the impact in the misclassification. Moreover, it was applied the same QMITP ($\phi = 0.99$) in all investigated cases. Hence, it seems that the shape of the applied distribution is not important to assess the problem of misclassification, but exclusively the value of ϕ.

5 Final Remarks

The usual $\mathbf{M_1}$ methodology can be applied to control the specificity and minimize the cost, as for screening cases, and our proposed $\mathbf{M_2}$ methodology can be applied in order to control the sensitivity, therefore in epidemic cases. The optimum group size n depends on your purpose, since it can be greater if our main goal is to save resources (and accepting having a higher probability of misclassification) or lower if the main goal is to control the problem of misclassification (cases in which we should use a smaller dimension n to ensure a low probability of misclassification). Furthermore, only if individual test has good performance (low probability of misclassification) the compound tests should be applied. Hence, the proposed quality measure of the individual test performance ϕ can be used to identify those situations. Moreover, the distribution $\mathbf{D}_0$ and the prevalence rate p do not seem to have a major impact on Dorfman's methodology misclassification problem if the same ϕ value is provided. A high ϕ value ensures a high φ_{s_n} and φ_{e_n} in the new $\mathbf{M_2}$ methodology, although the RC remains quite high. The same high ϕ value only guarantees a high φ_{e_n} in the usual $\mathbf{M_1}$ methodology and, therefore, it must be used with caution. Nevertheless, in the $\mathbf{M_1}$ methodology, the RC rapidly decreases with n.

Acknowledgements Research partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal – FCT under the project PEst-OE/MAT/UI0006/2011.

References

1. Berger, T., Mandell, J.W., Subrahmanya, P.: Maximally efficient two-stage screening. Biometrics **56**, 833–840 (2000)
2. Bilder, C.R., Zang, B., Schaarschmidt, F., Tebbs, J.M.: binGroup: a package for group testing. The R Journal **2**, 56–60 (2010)
3. Boswell, M.T., Gore, S.D., Lovison, G., Patil, G.P.: Annotated bibliography of composite sampling, Part A: 1936–92. Environ. Ecol. Stat. **3**, 1–50 (1996)
4. Chen, C., Swallow, W.: Sensitivity analysis of variable-sized group testing and its related continuous models. Biometrical J. **37**, 173–181 (1995)
5. Dorfman, R.: The detection of defective members in large populations. Ann. Math. Statistics **14**, 436–440 (1943)
6. Finucan, H.M.: The blood testing problem. Appl. Stat. – J. Roy. St. C **13**, 43–50 (1964).
7. Garner, F.C., Stapanian, M.A., Yfantis, E.A., Williams, L.R.: Probability estimation with sample compositing techniques. J. Off. Stat. **5**, 365–374 (1989)
8. Gastwirth, J.L.: The efficiency of pooling in the detection of rare mutations. Am. J. Hum. Genet. **67**, 1036–1039 (2000)

9. Hung, M., Swallow, W.: Robustness of group testing in the estimation of proportions. Biometrics **55**, 231–237 (1999)
10. Hughes-Oliver, J.M.: Pooling experiments for blood screening and drug discovery, Screening - Methods for Experimentation in Industry, Drug Discovery, and Genetics, Springer, 48–68 (2006)
11. Hwang, F.K.: Group testing with a dilution effect. Biometrika **63**, 671–673 (1976)
12. Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D., Pilcher, C.: Comparison of group testing algorithms for case identification in the presence of testing errors. Biometrics **63**, 1152–1163 (2007)
13. Johnson, N.L., Kotz, S., and Wu, X.: Inspection errors for attributes in quality control, Chapman and Hall (1991)
14. Lancaster, V.A., Keller-McNulty, S.: A review of composite sampling methods. J. Am. Stat. Assoc. **93**, 1216–1230 (1998)
15. Litvak, E., Tu, X.M., Pagano, M.: Screening for the presence of a disease by pooling sera samples. J. Am. Stat. Assoc. **89**, 424–434 (1994)
16. Liu, S.C., Chiang, K.S., Lin, C.H., Chung, W.C., Lin, S.H., Yang, T.C.: Cost analysis in choosing group size when group testing for potato virus Y in the presence of classification errors. Ann. Appl. Biology **159**, 491–502 (2011)
17. Loyer, M.W.: Bad probability, good statistics, and group testing for binomial estimation. Am. Stat. **37**, 57–59 (1983)
18. Martins, J.P., Santos, R., Sousa, R.: Testing the maximum by the mean in quantitative group tests. Studies in Theoretical and Applied Statistics, Selected Papers of the Statistical Societies (in print)
19. Phatarfod, R.M., Sudbury, A.: The use of a square array scheme in blood testing. Statistics in Medicine **13**, 2337–2343 (1994)
20. Roederer, M. and Koup, R A. Optimized determination of T cell epitope responses. Journal of Immunological Methods **274**, 221–228 (2003)
21. Santos, R., Pestana, D., Martins, J.P.: Extensions of Dorfman's Theory. In Oliveira, P.E. et al. (eds.): Studies in Theoretical and Applied Statistics, Recent Developments in Modeling and Applications in Statistics, 179–189, Springer (2013)
22. Sobel, M., Elashoff, R.: Group testing with a new goal, estimation. Biometrika **62**, 181–193 (1975)
23. Sobel, M., Groll, P.A.: Group testing to eliminate efficiently all defectives in a binomial sample. Bell Syst. Tech. J. **38**, 1179–1252 (1959)
24. Sterret, A.: On the detection of defective members of large populations. Ann. Math. Statistics **28**, 1033–1036 (1957)
25. Tu, X.M., Litvak, E., Pagano, M.: Studies of AIDS and HIV surveillance, Screening tests: Can we get more by doing less?. Stat. Med. **13**, 1905–1919 (1994)
26. Tu, X.M., Litvak, E., Pagano, M.: On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. Biometrika **82**(2), 287–297 (1995)
27. Wein, L.M., Zenios, S.A.: Pooled testing for HIV screening: capturing the dilution effect. Oper. Res. **44**, 543–569 (1996)
28. Woodbury, C.P., Fitzloff, J.F., Vincent, S.S.: Sample multiplexing for greater throughput in HPLC and related methods. Analytical Chemistry **67**, 885–890 (1995)
29. Zenios, S., Wein, L.: Pooled testing for HIV prevalence estimation exploiting the dilution effect. Stat. Med. **17**, 1447–1467 (1998)