Accuracy of the H_2S test: a systematic review of the influence of bacterial density and sample volume

Hong Yang, Jim A. Wright, Robert E. S. Bain, Steve Pedley, John Elliott and Stephen W. Gundry

ABSTRACT

The presence/absence hydrogen sulphide test (P/A H_2S) is widely used as a low-cost alternative faecal indicator test in remote and resource-poor settings. The aim of the paper is to assess how bacterial density and sample volume affect its accuracy. Based on a systematic search, we identified studies that tested water samples ($n = 2{,}034$) using both the P/A H_2S test and recognised tests for thermotolerant coliforms (TTC) or *Escherichia coli*. We calculated P/A H_2S test specificity and sensitivity against a range of TTC and *E. coli* densities. For two studies, we compared this with sensitivity and specificity estimates for simulated 100 and 20 ml presence/absence tests. For most of the 19 included studies, as the threshold used to define contamination increased from 1 to 100 cfu/100 ml, P/A H_2S test sensitivity increased but specificity decreased. Similarly, the simulation indicated that increasing test volumes from 20 to 100 ml increased sensitivity but reduced specificity. There was potential for bias, for example from lack of blinding during test interpretation, in most of the studies reviewed. In assessing the P/A H_2S test as an alternative to standard methods, careful consideration of likely indicator bacteria levels and sample volume is required.

Key words | *E. coli*, H_2S test, sensitivity, specificity, thermotolerant coliforms, water quality

Hong Yang (corresponding author)
Jim A. Wright
Geography and Environment,
University of Southampton,
University Road,
Southampton SO17 1BJ,
UK
E-mail: *hongyanghy@gmail.com*

Robert E. S. Bain
Stephen W. Gundry
Water and Health Research Centre,
University of Bristol,
Woodland Road,
Bristol BS8 1UB,
UK

Steve Pedley
John Elliott
Robens Centre for Public and Environmental Health,
University of Surrey,
Guildford,
Surrey GU2 7XH,
UK

INTRODUCTION

Diarrhoeal disease due to inadequate water, sanitation and hygiene caused 2.2 million deaths and 76.3 million Disability Adjusted Life Years (DALYs) in 2000 (Prüss *et al.* 2002). Microbial contamination is the most common and widespread health risk associated with drinking water (Yang *et al.* 2013) and this risk can be managed through water quality monitoring and remediation of contaminated supplies (WHO 2008). In many remote and developing country settings, monitoring for faecal contamination of drinking water is limited by the lack of laboratory facilities, finance and trained staff. Consequently, the hydrogen sulphide (H_2S) method was introduced by Manja *et al.* (1982) as a low-cost field test to detect faecal pollution of water in such settings. This method has been widely used as a presence/absence (P/A) test in developing countries and remote areas and has also been recommended by UNICEF (2007). The test has also been implemented as a multiple tube method (Roser *et al.* 2005; McMahan *et al.* 2012), but generally is used in P/A form to reduce both cost and complexity of use. Following Manja's original formulation, it is often used with a sample volume of 20 ml rather than the standard 100 ml, presumably to reduce cost per sample and for convenience. Various studies have determined its diagnostic accuracy relative to standard laboratory-based methods. A review has also summarised the performance of the H_2S method and its potential limitations for drinking water monitoring (Sobsey & Pfaender 2003). More recently, we conducted a systematic review and meta-analysis (Wright *et al.* 2012) of the diagnostic accuracy of the P/A H_2S test versus tests for thermotolerant coliforms (TTC) or *Escherichia coli*. We used a TTC or *E. coli* density of 1 cfu/100 ml to define contamination wherever possible.

The use of the P/A H_2S test raises a number of questions. First, there are situations where it may be desirable to detect a

doi: 10.2166/wh.2013.225

level of contamination higher than 1 cfu/100 ml. For example, WHO (2011) defines 'low risk' contamination (1–10 cfu/100 ml), 10–100 cfu/100 ml as intermediate risk, and 100–1,000 cfu/100 ml as high risk, with >1,000 cfu/100 ml being 'very high risk'. Similarly, for wastewater and environmental waters, guideline values are often higher than 1 cfu/100 ml (EU 2006; WHO 2006). How does the P/A H_2S test compare with standard indicator bacteria tests at these higher levels of contamination? Second, what effect does the choice of a 20 ml rather than 100 ml volume have on test results?

Looking at the test's ability to detect samples contaminated with greater densities of indicator bacteria (e.g., above 10 or 100 cfu/100 ml rather than ≥1 cfu/100 ml), three factors are likely to determine the diagnostic accuracy of the P/A H_2S test:

1. *Bacterial densities*: When bacterial densities are low, two 100 ml samples will be less likely to both test positive than if densities are higher. This is because of the inherent statistical variation in bacterial densities between one sample and the next. This is true for any microbiological test, not just the P/A H_2S test.
2. *Sample volume*: The P/A H_2S test is often implemented using a sample volume of 20 ml rather than the standard 100 ml, making it more difficult to detect low densities of indicator bacteria, particularly below 5 cfu/100 ml.
3. *Test procedures*: The P/A H_2S test detects a different group of indicator bacteria and incubation procedures vary.

In this paper, we examine the contribution of these three factors to P/A H_2S test accuracy by drawing on the results of a literature review, looking also at its ability to detect contamination not just above 1 cfu/100 ml, but other thresholds too (e.g., 10 or 100 cfu/100 ml). Such an analysis is very difficult when based on aggregate summary statistics in study reports comparing different tests. In contrast to previous quantitative reviews of water quality sample data, we therefore analyse individual sample data (ISD) rather than results that have been aggregated across an entire study or type of water source within a study. We also examine the effects of adopting a volume of 20 ml rather than 100 ml via simulation modelling. Two questions are addressed in this study. (1) How does indicator bacteria density influence the diagnostic accuracy of the P/A H_2S test? (2) How does sample volume influence the diagnostic accuracy of the P/A H_2S test?

METHODS

Strategy for literature search and data extraction

A copy of the review protocol is available from the authors; this paper reports on an analysis not foreseen in our original protocol. Eligible study characteristics, search strategy for identifying and procedures for characterising relevant literature have been described elsewhere (Wright *et al.* 2012). In brief, eligible studies were those that simultaneously tested drinking and surface water samples using both the P/A H_2S test and recognised tests for TTC or *E. coli*. We searched titles and abstracts from relevant bibliographic and grey literature databases using both English and Chinese (e.g., Microbiology Abstracts A and B; Water Resources Abstracts, Pollution Abstracts, Conference Papers Index, Web of Knowledge, Google Scholar, Compendex, GeoBase, Water Resources Worldwide, Medline, IndMed, British Library for Development Studies, Library of Congress Online Catalog, British Library Integrated Catalogue and WorldCat). For the full list of the searched database, see Table S1 in Wright *et al.* (2012), using terms for the H_2S test (e.g., 'H_2S', or 'hydrogen sulphide', or 'hydrogen sulfide', or 'pathoscreen', or 'Manja') with terms for domestic water samples (e.g., 'water', or 'environmental samples') and terms for indicator bacteria (e.g., 'thermotolerant', or 'faecal', or 'fecal', or 'coliform', or '*E. coli*') with a final search date of July 2010. References were also traced to and from included studies, from a relevant review by Sobsey & Pfaender (2003) and to the original paper describing the H_2S test (Manja *et al.* 1982). Where the full text of a paper was unavailable, we contacted one or more authors. Papers were then independently screened and characterised by two researchers (HY, JAW), with individual sample results being recorded independently in spreadsheets. Other characteristics recorded, including study quality criteria (Whiting *et al.* 2003) are described in Wright *et al.* (2012). Disagreements were resolved by consensus or referral to a third team member (SP, JE).

We included in our meta-analysis studies that reported individual samples and quantified TTC or *E. coli* densities.

We also included studies that reported aggregate results for groups of samples, but provided breakdowns of H_2S positive and negative samples for different TTC or *E. coli* bandings (e.g., 1–10, 10–100 and >100 cfu/100 ml). For samples with TTC or *E. coli* densities that were too numerous to count, we calculated the mid-point of the logged upper limit of detection and a value ten times this upper limit. For non-detectable values for indicator bacteria, we calculated the mid-point of the log of 0.5 and the logged lower limit of detection (Costa 2010).

Meta-analysis

To avoid the risk of Simpson's Paradox (Borenstein *et al.* 2009), all analyses were undertaken at study level, rather than pooling data across studies. We then calculated the number of samples that fell in four categories using a range of threshold TTC/*E. coli* density values that varied from 1 to 10,000 cfu/100 ml:

- true positives (tp) where the P/A H_2S test was positive and the TTC or *E. coli* density was above the threshold cfu/100 ml value;
- false positives (fp), where the P/A H_2S test was positive but the TTC or *E. coli* density was below the threshold value;
- false negatives (fn), where the P/A H_2S test was negative, but the TTC or *E. coli* density was above the threshold value;
- true negatives (tn), where the P/A H_2S test was negative and the TTC or *E. coli* density was below the threshold value.

For each threshold value of cfu/100 ml, we calculated two measures of H_2S test diagnostic accuracy from these counts, sensitivity and specificity. Sensitivity [$=tp/(tp+fn)$] is the proportion of water samples contaminated (above a threshold indicator bacteria density) that are correctly identified by the H_2S method. Specificity [$=tn/(tn+fp)$] is the proportion of uncontaminated water samples (below a given indicator bacteria threshold density) that are correctly identified by the H_2S method (Altman & Bland 1994). Sensitivity, the ability to correctly identify contaminated water, is particularly important in order to identify possible sources of faecal contamination. Specificity is also important as too many 'false alarms' waste limited resources. A test needs to be both sensitive and specific.

We tested for differences in sensitivity between the subset of studies included in this analysis and all of the studies included in our original review (Clarke & Stewart 2008) using a *t*-test in Stata version 11. A meta-regression was also conducted to test for differences in sensitivity and specificity between studies with and without individual sample results. For each study, we plotted sensitivity and specificity against the threshold cfu/100 ml value for TTC/*E. coli* that was used to define contaminated samples. We also plotted these values for studies that presented aggregate results of H_2S positive and negative samples broken down by multiple cfu/100 ml bands. Although we tested for heterogeneity and bias across all studies as reported in Wright *et al.* (2012), we were unable to test for heterogeneity among the subset of studies reporting individual sample results. Heterogeneity testing of such studies using Area Under the Receiver Operating Characteristic (ROC) curve was not possible because study results occupied different areas of ROC space.

Simulation

A small subset of the sample data from the meta-analysis was used in a simulation to evaluate the influence of different sample volumes (20 and 100 ml) and inherent statistical variation in bacterial densities on P/A H_2S diagnostic accuracy in more detail. Individual sample results were included as simulation input data provided they were derived using clearly documented methods for *E. coli* and TTC enumeration and had bacterial densities within the range of detection. Only 10 or 20 ml P/A H_2S samples were included from studies that had at least 20 such sample results. The simulations were performed using MATLAB® version R2011a.

For each sample result the bacterial density as measured by the quantitative device was assumed to be the underlying density of organisms in the source. Using the Poisson sampling model (Cochran 1950), random numbers were used for each of these densities to yield simulated results for paired idealised 10 or 20 ml P/A and quantitative devices. For simplicity, the membrane filtration method has been assumed to determine the exact number of bacteria in a 100 ml sample volume. The sensitivity and specificity per study were calculated based on these results, varying the minimum bacterial density that defined a contaminated sample from 1 to 10,000 cfu/100 ml. The above procedure was repeated a number of times (10,000) allowing mean values to be calculated for each threshold. The entire simulation was then repeated using an idealised P/A test with a volume of 100 ml rather than 10 or 20 ml.

In this way, for a small subset of studies, we were able to compare the concordance between the following:

1. An idealised quantitative test of 100 ml with an idealised P/A test of 100 ml (with differences largely due to the inherent statistical variation in bacteria counts between samples and consequent uncertainty in enumeration methods).
2. An idealised quantitative test of 100 ml with an idealised P/A test of 10 or 20 ml (with differences due to the inherent statistical variation in bacteria counts between samples and the two different volumes used).
3. The actual quantitative and 10 or 20 ml P/A H_2S test results from the field (with differences due to the inherent statistical variation in bacteria counts between samples, the two different volumes used, and the difference in media and target indicator organism).

RESULTS

Studies included in the meta-analysis

As described for our earlier systematic review (Wright *et al.* 2012), 51 studies were included initially (as shown by Box A, Figure 1). Five studies were included that reported aggregate counts of true positives, false positives, true negatives and false negatives for the P/A H_2S method based on different indicator organism bandings. In total, 502 TTC samples and 740 *E. coli* samples were simultaneously tested for H_2S-producing bacteria in these five studies (Table 1; Box F, Figure 1). Of the remaining 46 studies, 26 were excluded that only reported sample results in aggregate (Box B, Figure 1). A further four studies were excluded that reported individual sample results but used P/A methods to detect indicator bacteria (TTC or *E. coli*) and were therefore not amenable to analysis (Box C, Figure 1). Two further studies reporting ISD were also excluded because their P/A H_2S test results were all negative (Cervantes 2003) or positive (Mosley & Sharp 2005), preventing calculation of sensitivity or specificity (Box D, Figure 1). This left 14 included studies that reported test results for individual water samples. Two of these studies reported both TTC and *E. coli* results (Lukacs 2002; Nikaeen *et al.* 2010). In total, 520 TTC samples and 276 *E. coli* samples tested for H_2S-producing bacteria were reported in these 14 studies. Because of the pronounced effect of extreme outliers on sensitivity and specificity, we also excluded one sample with a very high TTC density of 1,600 cfu/100 ml (Ratto *et al.* 1990) and two samples with high *E. coli* densities (>5,000 and 500 cfu/100 ml) (Mattelet 2006), which all tested negative for H_2S (Box E, Figure 1). Overall, 1,021 TTC and 1,014 *E. coli* test results with associated P/A H_2S measurements were included in the meta-analysis.

Meta-analysis

Based on the 14 studies with ISD (Table 1; Box G, Figure 1), the percentage of P/A H_2S positive samples was calculated for different indicator bacteria densities (Table 2). Study quality criteria are presented in Table S1 of the Supplementary material (available online at http://www.iwaponline.com/jwh/011/225.pdf). There were some potential study design issues that affected all the included studies. For example, it was unclear in all studies whether those interpreting P/A H_2S test results were aware of TTC/*E. coli* results and vice versa.

We also examined the impact on sensitivity and specificity of varying the threshold cfu/100 ml value at which a sample was classified as positive according to a TTC or *E. coli* test. The change in sensitivity and specificity with increasing threshold density of TTC and *E. coli* for the five studies with aggregate sample results and 14 studies with ISD are shown in Figure 2. For almost all studies, sensitivity increased as the TTC or *E. coli* density threshold increased (Figures 2(a) and 2(c)), except for a few studies where it fluctuated (Monjour *et al.* 1986; Lukacs 2002; Coulbert 2005). Almost all studies show declining specificity as the threshold density of indicator bacteria increased (Figures 2(b) and 2(d)).

We tested for differences in study-level sensitivity between the subset of 19 studies described above and all 51 studies included in our earlier review (Wright *et al.* 2012). A *t*-test showed no significant differences for levels of TTC or *E. coli* ($p = 0.54$ and 0.64, respectively), nor did

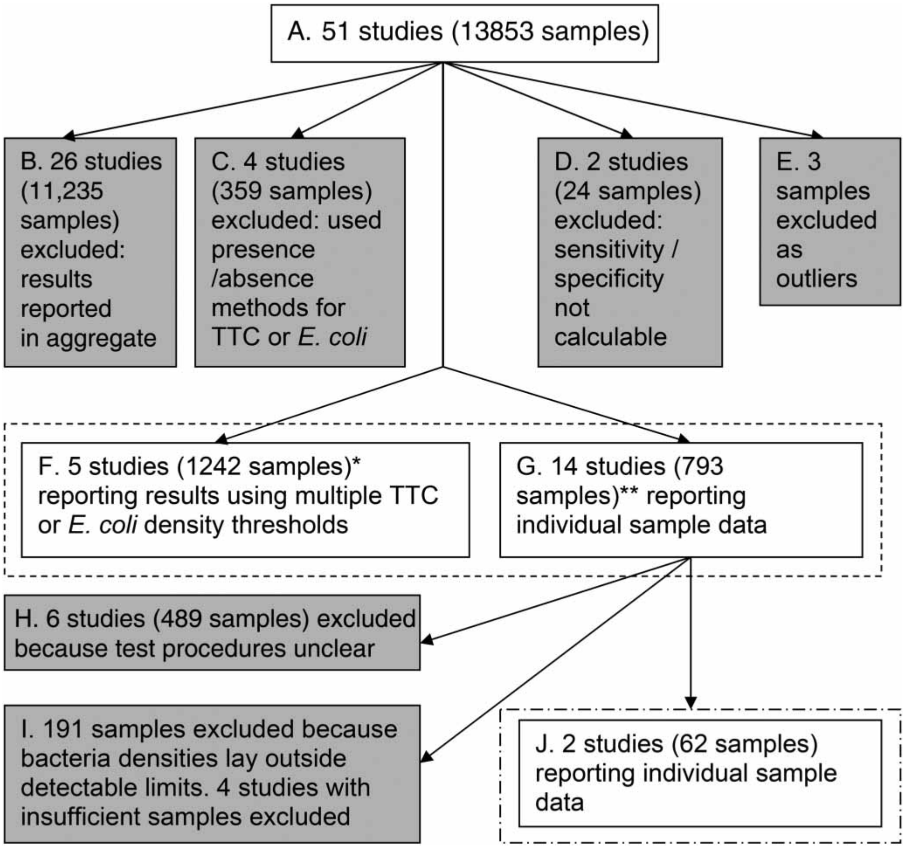
A. 51 studies (13853 samples)
B. 26 studies (11,235 samples) excluded: results reported in aggregate
C. 4 studies (359 samples) excluded: used presence /absence methods for TTC or *E. coli*
D. 2 studies (24 samples) excluded: sensitivity / specificity not calculable
E. 3 samples excluded as outliers
F. 5 studies (1242 samples)* reporting results using multiple TTC or *E. coli* density thresholds
G. 14 studies (793 samples)** reporting individual sample data
H. 6 studies (489 samples) excluded because test procedures unclear
I. 191 samples excluded because bacteria densities lay outside detectable limits. 4 studies with insufficient samples excluded
J. 2 studies (62 samples) reporting individual sample data

Figure 1 | Flow diagram, showing how studies and samples were selected for inclusion in the meta-analysis and subsequent simulation. —-, samples selected for inclusion in meta-analysis; -.-.-., samples selected for inclusion in simulation; *, 1,274 samples in total, but only 1,242 samples reporting results using multiple TTC or *E. coli* density thresholds; **, 875 samples in total, but only 793 reporting individual sample results without extreme outliers.

a meta-regression that tested for differences in sensitivity ($p = 0.114$) and specificity ($p = 0.636$) between studies that did and did not report individual sample results.

Simulation

Of the 14 included studies with published individual sample results described above, six studies were excluded because the descriptions of methods for enumerating TTC, *E. coli* and/or H_2S-producing bacteria were unclear (Box H, Figure 1). From the remaining eight studies, 191 samples were further excluded because they had TTC or *E. coli* densities outside detectable limits (Box I, Figure 1). Overall, 62 samples from two studies that contained sufficient sample numbers, information on test procedures and had TTC or *E. coli* densities within the detectable range were included in the simulation (Box J, Figure 1).

Simulation results have been plotted in Figure 3 alongside sensitivity and specificity calculated from the two studies described above; the results were plotted in ROC space, a form of graphical plot commonly used to illustrate the performance of a binary (P/A) diagnostic test. In particular, such plots are used to show the trade-off between sensitivity and specificity, since an increase in sensitivity is often at the expense of specificity, and vice versa. These graphs show how sensitivity and specificity change in two studies, Figure 3(a) Mosley *et al.* (2004) and Figure 3(b) Johnson (2007) for:

- a simulated, idealised quantitative test of 100 ml versus a simulated P/A test of 100 ml;
- a simulated quantitative test of 100 ml versus a simulated P/A test of 10 ml for Mosley *et al.* (2004) and 20 ml for Johnson (2007);

Table 1 | Summary of studies with individual sample data or reporting aggregated sample results using multiple threshold densities of indicator bacteria

			H_2S test procedure			**Characteristics of recognised method**	
Reference	**Studies reporting ISD or multiple threshold**[a]	**Number of samples**	**Incubation period (h)**	**Incubation temperature (°C)**	**Sample volume (ml)**	**Indicator bacteria**	**Geometric mean (cfu/100 ml)**
Coulbert (2005)	ISD	17	48	35	100	*E. coli*	3.7
Grant & Ziel (1996)	ISD	14	24	30	100	TTC	
Hewison *et al.* (1988)	ISD	128	24, 30	Ambient temperature		TTC	2.05
Johnson (2007)	ISD	61	48	25–35	20	*E. coli*	6.41
Kromoredjo & Fujioka (1991)	ISD	50	12–15, 18–24	26–30		*E. coli*	4.28
Lukacs (2002)	ISD	52	72	37	10, 20	TTC	1.34
Lukacs (2002)	ISD	15	72	37	10, 20	*E. coli*	3.03
Mattelet (2006)	ISD	46	48	35	20	*E. coli*	6.03
Monjour *et al.* (1986)	ISD	58	18–24	37	20	TTC	0.25
Mosley *et al.* (2004)	ISD	55	36	25–30	10	TTC	2.83
Nikaeen *et al.* (2010)	ISD	35	48	37	60, 100, 120	TTC	5.7
Nikaeen *et al.* (2010)	ISD	35	48	37	60, 100, 120	*E. coli*	2.15
Okioga (2007)	ISD	30	24, 48	24–35	20	*E. coli*	0.13
Peletz (2006)	ISD	22	24, 48	35	20	*E. coli*	5.71
Ratto *et al.* (1989)	ISD	20	12–18	22, 35	20	TTC	0.73
Ratto *et al.* (1990)	ISD	158		22, 35		TTC	0.86
Desmarchelier *et al.* (1992)	Multiple	358	18, 30	37, 38	20	*E. coli*	0.9
Genthe & Franck (1999)	Multiple	413	48	22, 35	20	TTC	8.59
Gupta *et al.* (2008)	Multiple	382	24, 48, 72	25	20	*E. coli*	17.65
Singh *et al.* (1985)	Multiple	35	12–18	37	20	TTC	38.74
Vasudevan & Tandon (2008)	Multiple	54	48	Ambient temperature	20	TTC	6.46

[a]Studies reporting individual sample data (ISD) or reporting samples using multiple indicator bacterial density thresholds.

- actual quantitative tests of 100 ml versus 20 ml P/A H_2S tests.

Comparing the simulated P/A tests of 100 ml (◊) and 10 or 20 ml (□), the simulation suggests a smaller volume lowers the sensitivity of a P/A test, but raises its specificity; this effect is greater in the Mosley *et al.* study where contamination levels were low. Comparing the actual H_2S test (▲) and simulated P/A of the same volume (□), the P/A H_2S test has much lower sensitivity, but somewhat higher specificity; an effect that is greater in the Johnson study where levels of *E. coli* were high. For equivalent values of specificity, in places the observed P/A H_2S test data show somewhat lower values for sensitivity compared with the 10 or 20 ml simulated P/A test.

DISCUSSION

Summary and implications for field testing

With respect to the diagnostic accuracy of the P/A H_2S test, we suggested three reasons for discrepancies between P/A

Table 2 | Percentages of H_2S positive samples by level of contamination with thermotolerant coliform (TTC) or *E. coli* (EC) (numbers in brackets are the total number of samples per level of TTC/EC contamination)

		Density of indicator bacteria (cfu/100 ml)			
Study	**Indicator bacteria**	**< 1**	**1–9.99**	**10–99.99**	**> 100**
Coulbert (2005)	EC	44% (9)	75% (8)		
Desmarchelier *et al.* (1992)	EC	15% (309)	57% (46)	81% (32)	
Genthe & Franck (1999)	TTC	22% (190)	71% (51)	92% (36)	100% (138)
Grant & Ziel (1996)	TTC	25% (4)	50% (2)	100% (6)	100% (2)
Gupta *et al.* (2008)	EC	3% (219)	7% (15)	27% (15)	81% (133)
Hewison *et al.* (1988)	TTC	30% (67)[a]	54% (46)[b]	73% (15)	
Johnson (2007)	EC	25% (32)	50% (4)	75% (4)	100% (21)
Kromoredjo & Fujioka (1991)	EC	4% (25)[a]	60% (5)[c]	100% (20)[d]	
Lukacs (2002)	TTC	15% (39)	44% (9)	50% (4)	100% (5)
Lukacs (2002)	EC	18% (11)	0% (1)	57% (7)	100% (1)
Mattelet (2006)	EC	45% (31)	50% (2)	0% (1)	75% (12)
Monjour *et al.* (1986)	TTC	15% (47)	22% (9)	50% (2)	
Mosley *et al.* (2004)	TTC	17% (23)	78% (18)	91% (11)	100% (3)
Nikaeen *et al.* (2010)	TTC	33% (9)[e]	94% (18)[f]	100% (8)[g]	
Nikaeen *et al.* (2010)	EC	57% (14)[e]	94% (17)[f]	100% (4)[g]	
Okioga (2007)	EC	11% (28)	100% (1)	100% (1)	
Peletz (2006)	EC	25% (8)	25% (4)	60% (5)	100% (5)
Ratto *et al.* (1988)	TTC	6% (145)[h]	80% (10)[i]	100% (3)	0% (1)
Ratto *et al.* (1989)	TTC	23% (13)	80% (5)	100% (2)	
Singh *et al.* (1985)	TTC	0% (2)	0% (9)	64% (11)	100% (13)
Vasudevan & Tandon (2008)	TTC	13% (15)	43% (14)	76% (25)	

For some studies, different TTC/EC contamination intervals were used because of variation in microbiological procedures and reporting of results:

[a] < 2.2 cfu/100 ml.

[b] 2.2–9.9 cfu/100 ml.

[c] 2.2–16 cfu/100 ml.

[d] > 16 cfu/100 ml.

[e] < 1.1 cfu/100 ml.

[f] 1.1–23 cfu/100 ml.

[g] > 23 cfu/100 ml.

[h] < 2 cfu/100 ml.

[i] 2–9.9 cfu/100 ml.

H_2S test results and those from standard laboratory methods, namely:

1. the inherent underlying statistical uncertainty from bacterial distributions in drinking water;
2. the frequent choice of 20 ml as a P/A H_2S test volume;
3. the difference in target indicator organisms and testing procedures.

As shown in Figure 3, each of these factors appears to reduce the overall degree of consistency between two sets of test results, measured using sensitivity and specificity. As the points representing an idealised 100 ml P/A (◊) in Figure 3 show, there is some discrepancy between the simulated results and a 100 ml quantitative test, as a result of inherent statistical uncertainty in bacterial sampling distributions.

This pattern is further affected by a reduction in P/A sample volume to 20 ml (Figure 3), which reduces sensitivity and increases specificity. This is consistent with a comparative study of different P/A H_2S volumes, which indicated that a 20 ml volume increased the percentage of H_2S positive samples relative to 10 ml (Roser *et al.* 2005). At the

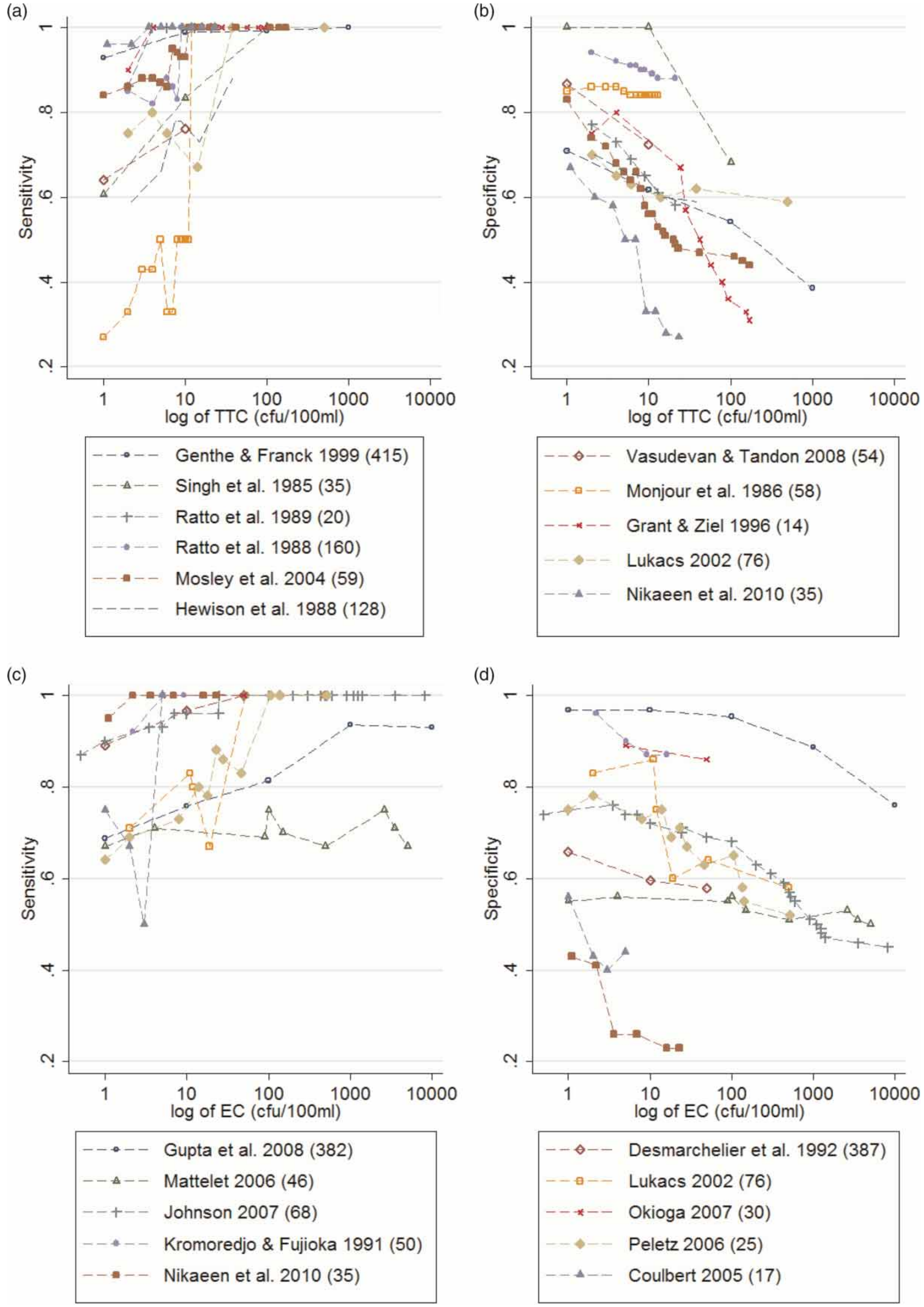

Figure 2 | Variation in sensitivity and specificity of the P/A H_2S method relative to varying threshold densities of thermotolerant coliforms and *E. coli* (numbers of samples per study in brackets). (a) Sensitivity of thermotolerant coliforms; (b) specificity of thermotolerant coliforms; (c) sensitivity of *E. coli*; (d) specificity of *E. coli*. A full colour version of this figure is available online at http://www.iwaponline.com/jwh/toc.htm.

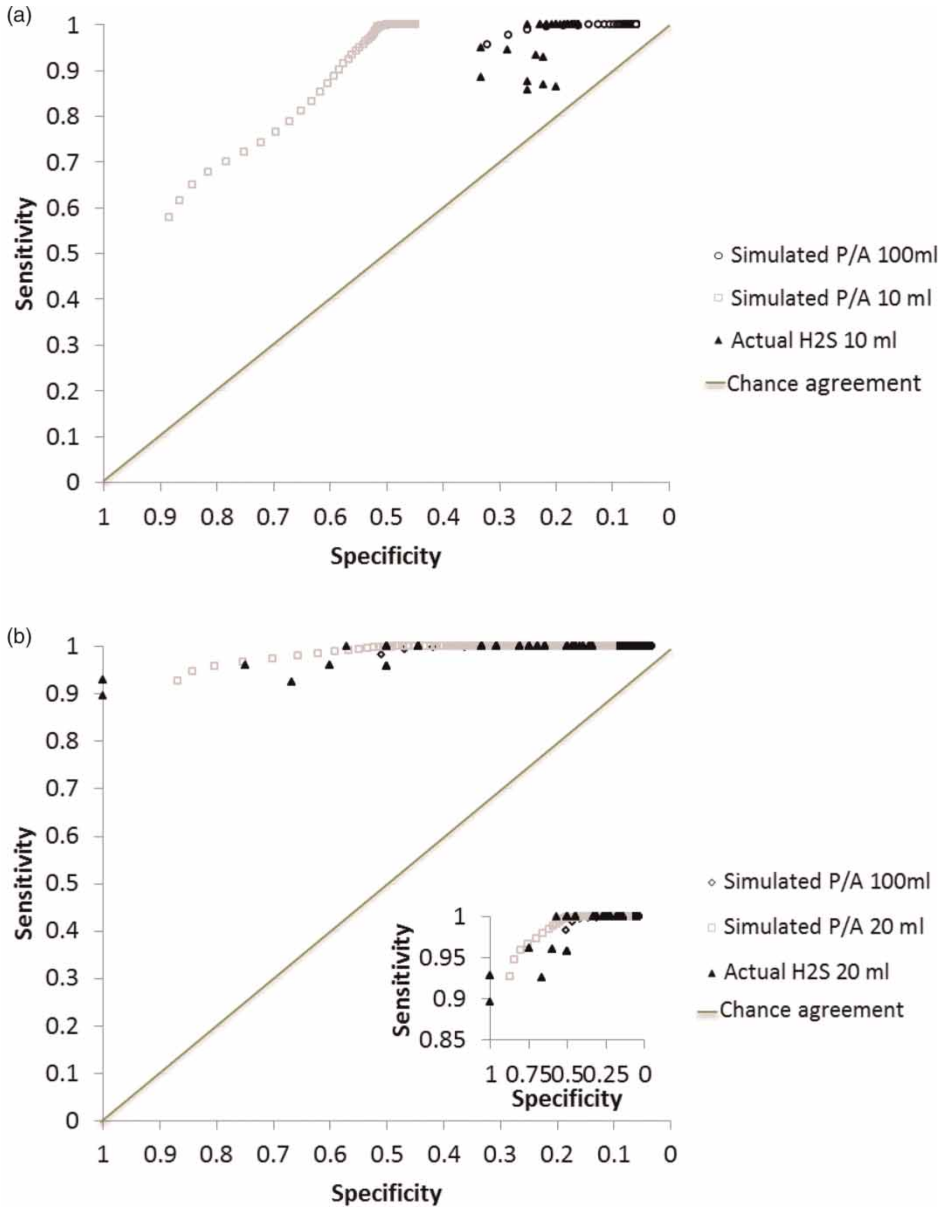

Figure 3 | Receiver Operating Characteristic space, showing sensitivity versus specificity relative to 100 ml quantitative tests for simulated presence/absence tests and actual P/A H_2S results from: (a) Mosley *et al.* (2004) (32 samples tested for TTC) and (b) Johnson (2007) (30 samples tested for *E. coli*).

same time, reducing the volume of a test reduces production costs. For example, Chuang *et al.* (2011) found that a 20 ml P/A H_2S test cost US$0.14, compared with a 100 ml P/A H_2S test which cost US$0.35. Lowering the test price per unit means that where budgets are limited, a greater number of tests can be conducted at the same cost. There is thus a trade-off between the unit cost of the test and its sensitivity. This trade-off should be considered when selecting a test volume for a given monitoring setting.

The sensitivity and specificity of the P/A H_2S test vary depending on the target level of contamination being detected (Figure 2). When the aim is to detect highly contaminated samples (e.g., with indicator bacteria >100 cfu/100 ml), the test's sensitivity is high but its specificity is low. When the aim is to detect lower levels of contamination (e.g., $\geq$1 cfu/100 ml), its sensitivity is lower but its specificity is higher. This trade-off between sensitivity and specificity at higher contamination levels is repeated across the many studies reviewed here. Thus, where sources are likely to be grossly contaminated (e.g., the Nicaraguan protected bucket wells reported by Sandiford *et al.* (1989)), the P/A H_2S test is likely to correctly identify a high proportion of such high risk sources, but a much lower proportion of low risk sources (e.g., the domiciliary piped connections reported by Sandiford *et al.* (1989)).

Thus, in designing a P/A test, there is a design space in which test volume, cost, sensitivity, specificity and the target level of contamination to be detected all interact in a complex manner. Following Manja's original formulation, many H_2S users implement the P/A H_2S test using a 20 ml volume rather than 100 ml (Table 1; Wright *et al.* 2012), which lowers its cost and sensitivity, but raises its specificity. Choosing a higher volume would alter these properties and so test users could, in theory, select a volume that best meets the needs of their specific situation.

Since H_2S producing bacteria are a different indicator organism group to TTC or *E. coli*, sensitivity at a given level of specificity is generally lower for the P/A H_2S test than for a simulated 20 ml P/A test. This effect is quantified in Figure 3. Previous studies and major reviews of the P/A H_2S test provide a number of explanations for the appearance of false positive results when compared with the standard tests for detection of faecal indicator bacteria (e.g., Sobsey & Pfaender 2003). Non-enteric bacteria that can reduce sulphate are ubiquitous, but many of the genera that comprise this group have exacting growth requirements that will reduce the potential for them giving a positive result. However, Sobsey & Pfaender (2003) do list other bacteria, for example *Pseudomonas* spp., *Bacillus* spp. and *Proteus* spp., that can be non-enteric and will give a positive result with the P/A H_2S test. This group of organisms are likely to produce a positive result with the P/A H_2S test when the standard tests for other indicators are negative. Several authors have reported false negative results at low densities of *E. coli*, but there are also a few studies that directly report a negative result for the P/A H_2S test for samples containing a high density of indicator bacteria (e.g., Monjour *et al.* 1986), or where the same result can be inferred from the sensitivity of the test at high densities of indicator bacteria (Gupta *et al.* 2008); these results had a pronounced effect on the sensitivity and specificity trends observed in Figure 2. The reason for these occasional results is not clear, and neither is it discussed by the authors. It is possible, although very unlikely at the levels of contamination being reported, that the contamination has a restricted range of indicator bacteria which excludes species that can reduce sulphate. There may be certain characteristics of the water that inhibit, or appear to inhibit the P/A H_2S test; once again, this seems unlikely. False negative results when indicator densities are high are a cause for concern and more research is needed to understand the conditions that can lead to this outcome.

Strengths and limitations

It is becoming increasingly common in clinical medicine to use data relating to individual patients in meta-analysis. The potential additional benefits of using individual patient data (IPD) rather than aggregated data on groups of patients include: increased statistical power, more flexible analysis of sample subgroups, more flexible analysis of outcomes and greater opportunities for data checking and correction (Clarke & Stewart 2008). To the best of our knowledge, the present systematic review is the first attempt to conduct a systematic review based on ISD rather than aggregated results for a water quality test. Using ISD enabled us to systematically vary the threshold TTC or *E. coli* density, scrutinise underlying raw data and explore effects of faecal bacteria density on the diagnostic accuracy of P/A H_2S test. This highlights the benefits of drawing on techniques used in other fields.

Various limitations were noted in our earlier analysis (Wright *et al.* 2012), particularly potential for bias in the studies reviewed. For example, in all studies, it was unclear whether those interpreting P/A H_2S results were aware of the results obtained from TTC/*E. coli* tests and vice versa. Furthermore, here we were only able to analyse a small proportion of relevant water samples (1,021 for TTC and 1,014 for *E. coli*). While there was no evidence of a systematic difference in sensitivity between this subset of samples and the overall set of samples identified, the available ISD for *E. coli* were largely drawn from unpublished reports and theses, rather than peer-reviewed articles. Arguably, such data may be systematically different from those collected by published studies and therefore biased. This study compared the P/A H_2S test with standard methods, *E. coli* and TTC, which are themselves imperfect measures of faecal contamination (Gleeson & Gray 1997); their direct comparison with the P/A H_2S test can therefore understate the ability of the P/A H_2S test to identify waters containing organisms of likely faecal origin (McMahan *et al.* 2012). In comparing sensitivity and specificity results from the two simulations and the reported P/A H_2S test results, in theory it would have been possible to test for significant differences between the area under two ROC curves shown in Figure 3 (DeLong *et al.* 1988). However, given that the two sets of outputs occupied different areas of ROC space, it would have been problematic to calculate the area under the two curves in a consistent way.

There are also limitations to the simulation:

1. Samples with values outside the detectable range of contamination have been excluded. Inclusion is likely to improve agreement between the tests.
2. The simulation is based on a very small proportion of the samples identified through the earlier review (32 samples for TTC and 30 samples for *E. coli*).
3. The measured values of the TTC and *E. coli* quantitative tests were assumed to be the true underlying bacterial density in the source. However, these estimates of true bacterial density are uncertain and our analysis did not account for this uncertainty.
4. The assumed Poisson distribution has been questioned (Haas & Heller 1988). Over-dispersion would result in greater inconsistency between results and therefore lower both sensitivity and specificity.

CONCLUSIONS

This analysis suggests that P/A H_2S test performance (measured by sensitivity and specificity) varies with indicator bacteria levels. Samples with no detectable TTC or *E. coli* are sometimes H_2S positive, while many samples with 1–9.99 cfu/100 ml of TTC or *E. coli* are H_2S negative (Table 2). For more heavily contaminated samples (above 10 cfu/100 ml of TTC or *E. coli*) the P/A test result is usually positive.

Sample volume also affects P/A H_2S test performance. The simulation results suggest a test volume of 100 ml gives higher sensitivity, but lower specificity, compared to a 20 ml test. However, test volume influences the cost of consumables, with a 20 ml volume more than halving the cost per test compared to a 100 ml volume (Chuang *et al.* 2011).

Taken together, these factors suggest that in a situation where there is a need to detect low levels of contamination (1–9.99 cfu/100 ml), it may be advisable to invest in a smaller number of more expensive 100 ml tests, rather than more widespread use of cheaper 20 ml tests. In situations where there are likely to be higher levels of contamination, using a larger number of more affordable 20 ml tests may be appropriate. However, although in theory, sample volume can be optimised in this way to obtain a balance between the diagnostic performance and cost, practical considerations about operationalising monitoring may override any decision as to the most appropriate test procedures for a given situation.

We encourage researchers and practitioners to make ISD available for future analyses. Analysis of ISD can yield valuable insights into the relative performance of diagnostic tests. These insights are especially important for low-income countries where difficult trade-offs must be made between test cost and performance.

ACKNOWLEDGEMENTS

This research has been funded by the Bill & Melinda Gates Foundation, under a grant to the University of Bristol. The grant's purpose is to develop a low cost test of microbial water quality for use in developing countries. We wish to acknowledge valuable methodological comments by Prof. Paul Hunter, University of East Anglia, on an earlier version of this manuscript.

REFERENCES

Altman, D. G. & Bland, J. M. 1994 Diagnostic tests. 1: Sensitivity and specificity. *Brit. Med. J.* **308**, 1552.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. 2009 Simpon's paradox. In: *Introduction to Meta-analysis* (M. Borenstein, L. V. Hedges, J. P. T. Higgins & H. R.

Rothstein, eds). Wiley Online Library, Chichester, UK, pp. 303–305.

Cervantes, D. X. F. 2003 Feasibility of Semi-continuous Solar Disinfection System for Developing Countries at a Household Level. Master Thesis, MIT, Cambridge, MA.

Chuang, P., Trottier, S. & Murcott, S. 2011 Comparison and verification of four field-based microbiological tests: H_2S test, Easygel, Colilert, Petrifilm^{TM}. *J. Water Sanit. Hyg. Devel.* **1**, 68–85.

Clarke, M. J. & Stewart, L. 2008 Obtaining individual patient data from randomised controlled trials. In: *Systematic Reviews in Health Care: Meta-analysis in Context* (M. Egger, G. D. Smith & D. G. Altman, eds). BMJ Publishing Group, London, pp. 109–121.

Cochran, W. G. 1950 Estimation of bacterial densities by means of the 'Most Probable Number'. *Biometrics* **2**, 105–116.

Costa, J. 2010 *Calculating Geometric Means*. Buzzards Bay Natural Estuary Program. Available at: http://www.buzzardsbay.org/geomean.htm.

Coulbert, B. 2005 An Evaluation of Household Drinking Water Treatment Systems in Peru: The Table Filter and the Safe Water System. Master Thesis, MIT, Cambridge, MA.

DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. 1988 Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44** (3), 837–845.

Desmarchelier, P., Lew, A., Caique, W., Knight, S., Toodayan, W., Isa, A. R. & Barnes, A. 1992 An evaluation of the hydrogen sulfide water screening-test and coliform counts for water-quality assessment in rural Malaysia. *Trans. R. Soc. Trop. Med. Hyg.* **86**, 448–450.

EU 2006 Directive 2006/7/EC of the European Parliament and of the Council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EEC. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006L0007:EN:NOT.

Genthe, B. & Franck, M. 1999 A tool for addressing microbial water quality in small community water supplies: an H2S strip test. Water Research Commission Report No. 961/1/99, Pretoria, South Africa.

Gleeson, C. & Gray, N. F. 1997 *The Coliform Index and Waterborne Disease: Problems of Microbial Drinking Water Assessment*. E & FN Spon, London.

Grant, M. & Ziel, C. 1996 Evaluation of a simple screening test for faecal pollution in water. *J. Water Supply Res. Tech.* **45**, 13–18.

Gupta, S. K., Sheikh, M. A., Islam, M. S., Rahman, K. S., Jahan, N., Rahman, M. M., Hoekstra, R. M., Johnston, R., Ram, P. K. & Luby, S. 2008 Usefulness of the hydrogen sulfide test for assessment of water quality in Bangladesh. *J. Appl. Microbiol.* **104**, 388–395.

Haas, C. N. & Heller, B. 1988 Test of the validity of the Poisson assumption for analysis of most-probable-number results. *Appl. Environ. Microb.* **54**, 2996–3002.

Hewison, K., Mack, K. F. & Sivaborvorn, K. 1988 Evaluation of a hydrogen sulphide screening test. AIDAB, Thai-Australian Northeast Village Water Resource Project, report no 47, Melbourne, Australia.

Johnson, S. M. 2007 Health and Water Quality Monitoring of Pure Home Water's Ceramic Filter Dissemination in the Northern Region of Ghana. Master Thesis, MIT, Cambridge, MA.

Kromoredjo, P. & Fujioka, R. S. 1991 Evaluating three simple methods to assess the microbial quality of drinking-water in Indonesia. *Environ. Toxic. Water Qual.* **6**, 259–270.

Lukacs, H. A. 2002 From Design to Implementation: Innovative Slow sand Filtration for Use in Developing Countries. Master Thesis, MIT, Cambridge, MA.

Manja, K. S., Maurya, M. S. & Rao, K. M. 1982 A simple field test for the detection of faecal pollution in drinking water. *B. World Health Organization* **60**, 797–801.

Mattelet, C. 2006 Household Ceramic Water Filter Evaluation using Three Simple Low-cost Methods: Membrane Filtration, 3M Petrifilm and Hydrogen Sulfide Bacteria in Northern Region, Ghana. Master Thesis, MIT, Cambridge, MA.

McMahan, L., Grunden, A. M., Devine, A. A. & Sobsey, M. D. 2012 Evaluation of a quantitative H_2S MPN test for fecal microbes analysis of water using biochemical and molecular identification. *Water Res.* **46**, 1693–1704.

Monjour, L., Henry, P., Guillemin, F., Spinasse, A., Lagneaux, F., Alfred, C., Colin, J. J. & Gentilini, M. 1986 Comparative study of research methods to assess water pollution of the fecal origin in the rural Sudan-Western Sahara environment. *Bull. Soc. Pathol. Exot. Filiales* **79**, 549–556 (in French).

Mosley, L. M. & Sharp, D. S. 2005 The hydrogen sulphide (H2S) paper-strip test: a simple test for monitoring drinking water quality in the Pacific Islands. SOPAC Technical Report 373, Suva, Fiji Islands.

Mosley, L. M., Sharp, D. S. & Singh, S. 2004 Effects of a tropical cyclone on the drinking-water quality of a remote Pacific Island. *Disasters* **28**, 405–417.

Nikaeen, M., Izadi, M., Sabzali, A., Bina, B., Jonidi Jafari, N. A., Hatamzdeh, M. & Farrokhzadeh, H. 2010 The effects of incubation period and temperature on the hydrogen sulphide (H_2S) technique for detection of faecal contamination in water. *Afr. J. Environ. Sci. Technol.* **4**, 084–091.

Okioga, T. 2007 Water Quality and Business Aspects of Sachet-vended Water in Tamale, Ghana. Master Thesis, MIT, Cambridge, MA.

Peletz, R. L. 2006 Cross-sectional Epidemiological Study on Water and Sanitation Practices in the Northern Region of Ghana. Master Thesis, University of California, Berkeley, CA.

Prüss, A., Kay, D., Fewtrell, L. & Bartram, J. 2002 Estimating the burden of disease from water, sanitation, and hygiene at a global level. *Environ. Health Persp.* **110**, 537–542.

Ratto, A., Dutka, B. J., Vega, C., Lopez, C. & Elshaarawi, A. 1989 Potable water safety assessed by coliphage and bacterial tests. *Water Res.* **23**, 253–255.

Ratto, A., El-Shaarawi, A. H., Dutka, B. J., Lopez, C. & Vega, C. 1988 Coliphage association with coliform indicators: a case study in Peru. *Toxic. Assess.* **3**, 519–533.

Ratto, M. A., Lette, C. V., Lopez, C., Mantilla, H. & Apoloni, L. M. 1990 Evaluation of the coliphage procedure and the presence/absence test as simple, rapid, economical methods for screening potable water sources and potable water supplies in Peru. In: *Use of Simple, Inexpensive Microbial Water Quality Tests: Results of a Three-continent, Eight-country Research Project* (B. J. Dutka & A. H. El-Shaarawi, eds). International Development Research Centre, Ontario, Canada, pp. 61–78.

Roser, D. J., Ashbolt, N., Ho, G., Mathew, K., Nair, J., Ryken-Rapp, D. & Toze, S. 2005 Hydrogen sulphide production tests and the detection of groundwater faecal contamination by septic seepage. *Water Sci. Technol.* **51**, 291–300.

Sandiford, P., Gorter, A. C., Smith, G. D. & Pauw, J. P. 1989 Determinants of drinking water quality in rural Nicaragua. *Epidemiol. Infect.* **102**, 429–438.

Singh, M. P., Srivastata, R. K., Bachani, D. & Khare, K. K. 1985 Detection of faecal pollution of drinking water in rural areas: some observations. *Indian J. Prev. Soc. Med.* **16**, 12–15.

Sobsey, M. D. & Pfaender, F. K. 2003 *Evaluation of the H2S Method for Detection of Fecal Contamination of Drinking Water.* World Health Organization, Geneva.

UNICEF 2007 *UNICEF Handbook on Water Quality.* UNICEF, New York.

Vasudevan, P. & Tandon, M. 2008 Microbial quality of rainwater from roof surfaces. *J. Sci. Ind. Res.* **67**, 432–435.

Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Bossuyt, P. M. M. & Kleijnen, J. 2003 The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med. Res. Methodol.* **3**, 25.

WHO 2006 *WHO Guidelines for the Safe Use of Wastewater, Excreta and Greywater. Volume 1, Policy and Regulatory Aspects.* WHO, Geneva.

WHO 2008 *Guidelines for Drinking-water Quality Volume 1: Recommendations*, 3rd edition. World Health Organization, Geneva.

WHO 2011 *Guidelines for Drinking-Water Quality* (4th edition). WHO, Geneva.

Wright, J. A., Yang, H., Walker, K., Pedley, S., Elliott, J. & Gundry, S. W. 2012 The H_2S test versus standard indicator bacterial tests for faecal contamination of water: a systematic review and meta-analysis. *Trop. Med. Int. Health* **17**, 94–105.

Yang, H., Bain, R., Bartram, J., Gundry, S., Pedley, S. & Wright, J. 2013 Water safety and inequality in access to drinking-water between rich and poor households. *Environ. Sci. Technol.* **47**, 1222–1230.

First received 20 November 2012; accepted in revised form 5 March 2013. Available online 4 April 2013