BashTheBug: a crowd of volunteers reproducibly and accurately measure the minimum inhibitory concentrations of 13 antitubercular drugs from photographs of 96-well broth microdilution plates.

Philip W Fowler*[1], Carla Wright[1], Helen Spiers-Bowers[2], Tingting Zhu[3], Elisabeth ML Baeten[4], Sarah W Hoosdally[1], Ana Luíza Gibertoni Cruz[1], Aysha Roohi[1], Samaneh Kouchaki[3], Timothy M Walker[1], Timothy EA Peto[1], Grant Miller[2], Chris Lintott[2], David Clifton[3], Derrick W Crook[1], A Sarah Walker[1], and The CRyPTIC Consortium[5]

[1]Nuffield Department of Medicine, John Radcliffe Hospital, University of Oxford, Headley Way, Oxford, OX3 9DU, UK

[2]Zooniverse, Department of Physics, University of Oxford, Oxford, UK

[3]Institute of Biomedical Engineering, University of Oxford, UK

[4]Citizen Scientist, c/o Zooniverse, Department of Physics, University of Oxford, Oxford, UK

Abstract

Tuberculosis is a respiratory disease that is treatable with antibiotics. An increasing prevalence of resistance means that to ensure a good treatment outcome it is desirable to test the susceptibility of each infection to different antibiotics. Conventionally this is done by culturing a clinical sample and then exposing aliquots to a panel of antibiotics, thereby determining the minimum inhibitory concentration (MIC) of each drug. Using 96-well broth micro dilution plates with each well containing a lyophilised pre-determined amount of an antibiotic is a convenient and cost-effective way to measure the MICs of several drugs at once for a clinical sample. Although accurate, this is an expensive and slow process that requires highly-skilled and experienced laboratory scientists. Here we show that, through the BashTheBug project hosted on the Zooniverse citizen science platform, a crowd of volunteers can reproducibly and accurately determine the MICs for 13 drugs and that simply taking the median or mode of 11-17 independent classifications is sufficient. There is therefore a potential role for crowds to support (but not supplant) the role of experts in antibiotic susceptibility testing.

*To whom correspondence should be addressed: philip.fowler@ndm.ox.ac.uk, @philipwfowler

INTRODUCTION

Tuberculosis (TB) is a treatable (primarily) respiratory disease that caused illness in ten million people in 2019, with 1.4 million deaths[1]. Ordinarily this is more than any other single pathogen, however SARS-CoV-2 killed more people than TB in 2020 and is likely to do so again in 2021. Like all bacterial diseases treated with antibiotics, an increasing proportion of TB cases are resistant to one or more drugs.

Tackling this 'silent pandemic' will require action on several fronts, including the development of new antibiotics, better stewardship of existing antibiotics and much wider use of antibiotic susceptibility testing (AST) to guide prescribing decisions[2]. The prevailing AST paradigm is *culture-based*: a sample taken from the patient is grown and the pathogen identified. If required, further samples are cultured in the presence of different antibiotics and each test is inspected/measured to see which compounds inhibit the growth of the bacterium. A scientific laboratory, with an enhanced biosafety level and staffed by experienced and highly-trained laboratory scientists, is required to carry out such AST. Maintaining such laboratories with a cadre of expert scientists is expensive and hence they tend to be found only at larger hospitals and national public health agencies, even in high-income countries. This model, whilst effective, is practically and economically difficult to scale up which explains in part why most antibiotic prescribing decisions are still done without any AST data.

The key AST measurement that informs prescribing decisions is the *minimum inhibitory concentration* (MIC) of an antibiotic: this is the smallest measured concentration of an antibiotic that prevents growth of the pathogen. Conventionally, the sample is inoculated into an appropriate growth medium that contains the antibiotic at a range of concentrations, each of which is double that of the last. If the MIC is below a pre-determined threshold (cutoff) then the clinical isolate is classified as being susceptible to that drug and the attending clinician can have some confidence that it would be effective, should they choose to prescribe it. Historically it has been assumed has been that, since accuracy is paramount, only highly-trained and experienced laboratory scientists (experts), or more recently, extensively-validated automatic algorithms part of accredited AST devices, can measure MICs. In this paper we shall show that a crowd of volunteers, who have no microbiological training, can reproducibly and accurately determine the growth of *M. tuberculosis* (the causative agent of TB) on a 96-well plate and thence the MICs for 13 different antibiotics.

The BashTheBug citizen science project, which was launched in April 2017 on the Zooniverse platform, has two goals: (i) to help reduce MIC measurement error in the large dataset of $>$ 20,000 clinical *M. tuberculosis* isolates collected by the Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC) project and (ii) to provide a large dataset of classifications to train machine-learning models, thereby assessing their suitability. CRyPTIC is seeking to add to and therefore improve upon the existing catalogues that describe genetic variants that are associated with resistance to specific antituberculars[3–5] which would in turn accelerate the shift from culture-based to genetics-based AST which is faster and cheaper than the culture-based

alternatives[6] and is already well underway for tuberculosis[7]. Ultimately shifting to a *genetics-based* paradigm, where the susceptibility of a pathogen to an antibiotic is inferred from the genome of the pathogen, potentially offers a route to making pathogen diagnostics much more widely available.

METHODS

Plate design

The CRyPTIC project is collecting a large number ($> 20,000$) of clinical TB samples, each having its whole genome sequenced and the MIC of a panel of 14 antibiotics measured using a bespoke 96-well broth microdilution plate. This plate, called UKMYC5, is variant of the MYCOTB 96-well microdilution plate manufactured by Thermo Fisher and contains 14 anti-TB drugs. UKMYC5 includes two repurposed compounds (linezolid and clofazimine) and two new compounds (delamanid and bedaquiline). Since 96-well plates have 8 rows and 12 columns, fitting 14 drugs, alongside two positive control wells, onto the plate necessitated a complex design (Fig. S2).

Each of the antibiotics on the UKMYC5 plate is present at 5-8 concentrations, each double the previous. The smallest concentration that prevents growth of *M. tuberculosis* after two weeks of incubation is the minimum inhibitory concentration (MIC). Since *M. tuberculosis* is notoriously difficult to culture and inspect, relying on a reading taken by a single expert (laboratory scientist) would likely have led to sufficiently high levels of errors in the dataset to bedevil sophisticated analyses, such as genome-wide analysis studies.

Image dataset

Thirty one vials containing 19 external quality assessment (EQA) *M. tuberculosis* strains, including the reference strain H37Rv ATCC 27294[8], were sent to seven participating laboratories as described previously[9]. Since some labs only received a subset of the 31 vials (Table S1), a total of 447 plates were inoculated and then incubated for 3 weeks (Fig. 1A). Minimum inhibitory concentrations of the 14 drugs on the plate were measured after 7, 10, 14 and 21 days by two laboratory scientists using a Thermo Fisher Sensititre Vizion Digital MIC viewing system, a mirrored-box and a microscope. One or two photographs were also taken each time using the Vizion instrument (Fig. 1B).

A previous blinded validation study involving seven CRyPTIC laboratories showed that the UKMYC5 plate is reproducible and that it is optimal to read the plate using either a Thermo Fisher Vizion instrument or a mirrored-box after 14 days of incubation[9]. That study also showed that *para*-aminosalicylic acid (PAS) performed poorly and therefore this drug is excluded in all subsequent analyses. Each image was also processed and analysed by some bespoke software, the Automated Mycobacterial Growth Detection Algorithm (AMyGDA), that segmented each photograph, thereby providing a second independent MIC reading of all the drugs on each plate (Fig. 1B)[10].

Early internal tests using the Zooniverse platform showed that asking a volunteers to examine all 96 wells on a plate was too complex a task. We therefore observed a clinical microbiologist as she examined several photographs of UKMYC5 plates. Rather than considering the *absolute* growth in each well, she was constantly comparing the growth in the wells containing antibiotic back to the positive control wells and therefore judging what constituted

growth *relative* to how well the isolate had grown in the absence of drug. A suitable task was therefore to classify the growth in the wells for a single drug as long as the positive control wells are also provided. The AMyGDA software was therefore modified to composite such *drug images* (Fig. 1B).

Each UKMYC5 plate yielded 14 composite images, one for each drug. Throughout the following analysis we shall aggregate all the data from the different drugs on the UKMYC5 plate. To facilitate this we shall therefore consider the *dilution*, which is defined as the number of well in the drug image with the lowest antibiotic concentration which prevents bacterial growth, rather than the minimum inhibitory concentration. Following upload to the Zooniverse platform, the retirement threshold was set to 17 classifications, however some images attracted additional classifications with 191 images having ≥ 34 classifications whilst 83 have 100 or greater (Table S2, Fig. S3).

Analysis

The resulting classifications were downloaded from the Zooniverse platform, either by a web browser or using the panoptes-cli command line tool[11]. Two Python modules were written to parse, store, manipulate and graph this classification data. The first, `pyniverse`[12], is designed to be generic for Zooniverse projects whilst the second, `bashthebug`[13], inherits from the first and adds functionality specific to BashTheBug (Fig. 1C). Both are freely available to download and use. These Python modules output several Pandas[14] dataframes which were then indexed, filtered and joined to other dataframes containing the sample information and the MIC readings taken by the expert and the AMyGDA software. AMyGDA also measured the growth in the two positive control wells and this was also recorded in a dataframe. All subsequent analysis was performed using Python3 in a jupyter-notebook and all graphs were plotted using matplotlib.

Engagement

In addition to the Zooniverse project page, which contained background information, a tutorial and FAQs, we setup a blog[15] and a variety of social media channels, focussing mainly on Twitter (@bashthebug). These all used a professionally designed logo and typeface (Fig. 1C), allowing instantaneous recognition of the project, which is important since the Zooniverse platform hosts tens of projects at any one time, and to indirectly convey that this is a professional project and therefore of scientific and societal importance. Since the blog was launched in March 2017 we have written 71 posts, attracting 7,393 visitors who made 13,811 views. At the time of writing, the Twitter account, @bashthebug, has 393 followers and has tweeted 400 times. Finally, the volunteers interacted with one another as well as the project team via the BashTheBug Talk Boards on the Zooniverse platform. A total of 6,255 posts were made by 1,042 individuals on 4,834 topics. During the course of the project, one of our more experienced volunteers (EMLB) became a *de facto* moderator by answering so many of the questions posted ($>$500) which we recognised by giving her moderator status.

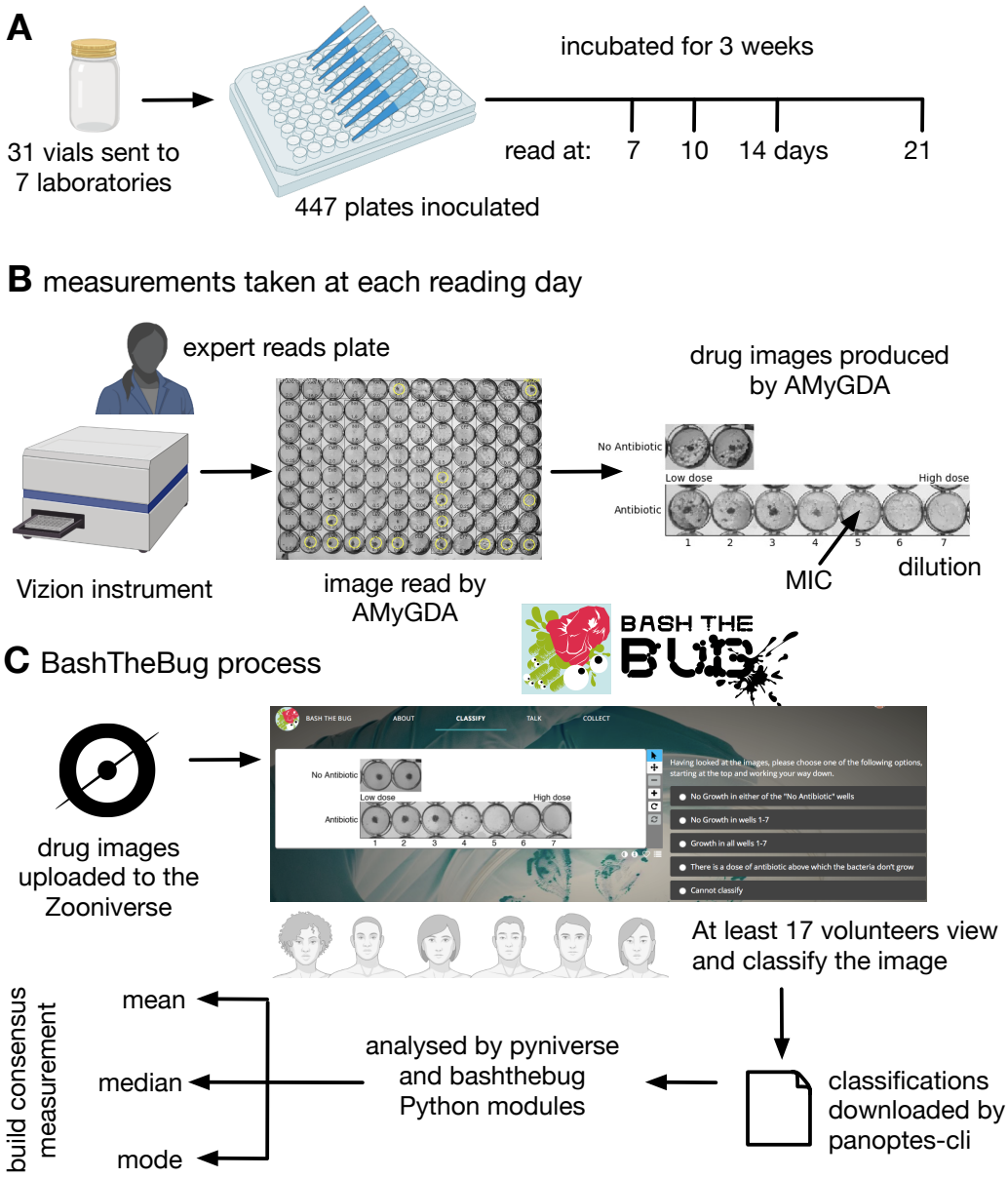

Figure 1: Each UKMYC5 plate was read by an Expert, by some software (AMyGDA) and by at least 17 citizen scientist volunteers via the BashTheBug project. (**A**) 447 UKMYC5 plates were prepared and read after 7, 10, 14 and 21 days incubation. (**B**) The minimum inhibitory concentrations (MIC) for the 14 drugs on each plate were read by an by Expert, using a Vizion instrument. The Vizion also took a photograph which was subsequently analysed by AMyGDA – this software then composited 14 drug images from each photograph, each containing an image of the two positive control wells. To allow data from different drugs to be aggregated, all MICs were converted to dilutions. (**C**) All drug images were then uploaded to the Zooniverse platform before being shown to volunteers through their web browser. Images were retired once they had been classified by 17 different volunteers. Classification data were downloaded and processed using two Python modules (pyniverse + bashthebug) before consensus measurements being built using different methods.

RESULTS

Project launch and progress

After a successful beta-test on 22 March 2017, BashTheBug was launched as a project on the Zooniverse citizen science platform on 8 April 2017. The launch was publicised on social media, mainly Twitter, and mentioned on several websites and the Zooniverse users were notified via email. By the end of the first week, 2,029 people had participated (of which 1,259 had usernames) classifying a total of 74,949 images – this includes the beta-test. The initial set of images were completed on 11 October 2017 and a second set was classified between 8 June 2020 and 15 November 2020 (Fig. 2A).

Volunteers

In total, 9,063 volunteers participated in classifying this dataset doing a total of 776,119 classifications (Fig. S1). The number of citizen scientists is an over-estimate since users who did not register with the Zooniverse (and therefore could not be identified through their unique username) but did more than one session will be counted multiple times. This is a mean of 85.6 classifications per volunteer, however this hides a large amount of variation in the number of classifications done by individual volunteers. Almost half of the volunteers (4,154) did ten or fewer classifications and 1,060 classified only a single image whilst the ten volunteers who participated the most did 103,569 classifications between them which is 13.3% of the total. The Gini-coefficient is a useful way to measure these unequal levels of participation, and for this dataset it is 0.85 (Fig. 2B).

Comparison to other Zooniverse projects

The activity within the first 100 days of launch has been used to benchmark and compare different Zooniverse projects from several academic disciplines[16]. A total of 381,964 classifications were done in the first hundred days after launch by a total of 6,237 users of which 3,733 were registered and so were unique. Several Zooniverse projects have attracted many more users and classifications, however, these are all ecology or astronomy projects which are the mainstay of the Zooniverse.

Since the number of classifications is heavily influenced by the difficulty of the task, it can be more illuminating to compare the Gini coefficients of different projects. Cox. *et al.*[17] measured a mean Gini coefficient across several Zooniverse projects of 0.81, whilst a later and more comprehensive study[16] demonstrated that Zooniverse projects had Gini coefficients in the range 0.54 to 0.94 with a mean of 0.80. They also suggested that biomedical projects had lower Gini coefficients, with a mean Gini coefficient of 0.67, however this was only based on three projects. BashTheBug attracted more users, completed more classifications and had a higher Gini coefficient than any of these three biomedical projects[16]. A more recent biomedical project, Etch-a-cell, that launched at a similar time to BashTheBug had a Gini coefficient of 0.83[18]. BashTheBug therefore has a higher than average level of participation inequality, having the 17th highest Gini coefficient out of 63 Zooniverse projects surveyed[16].

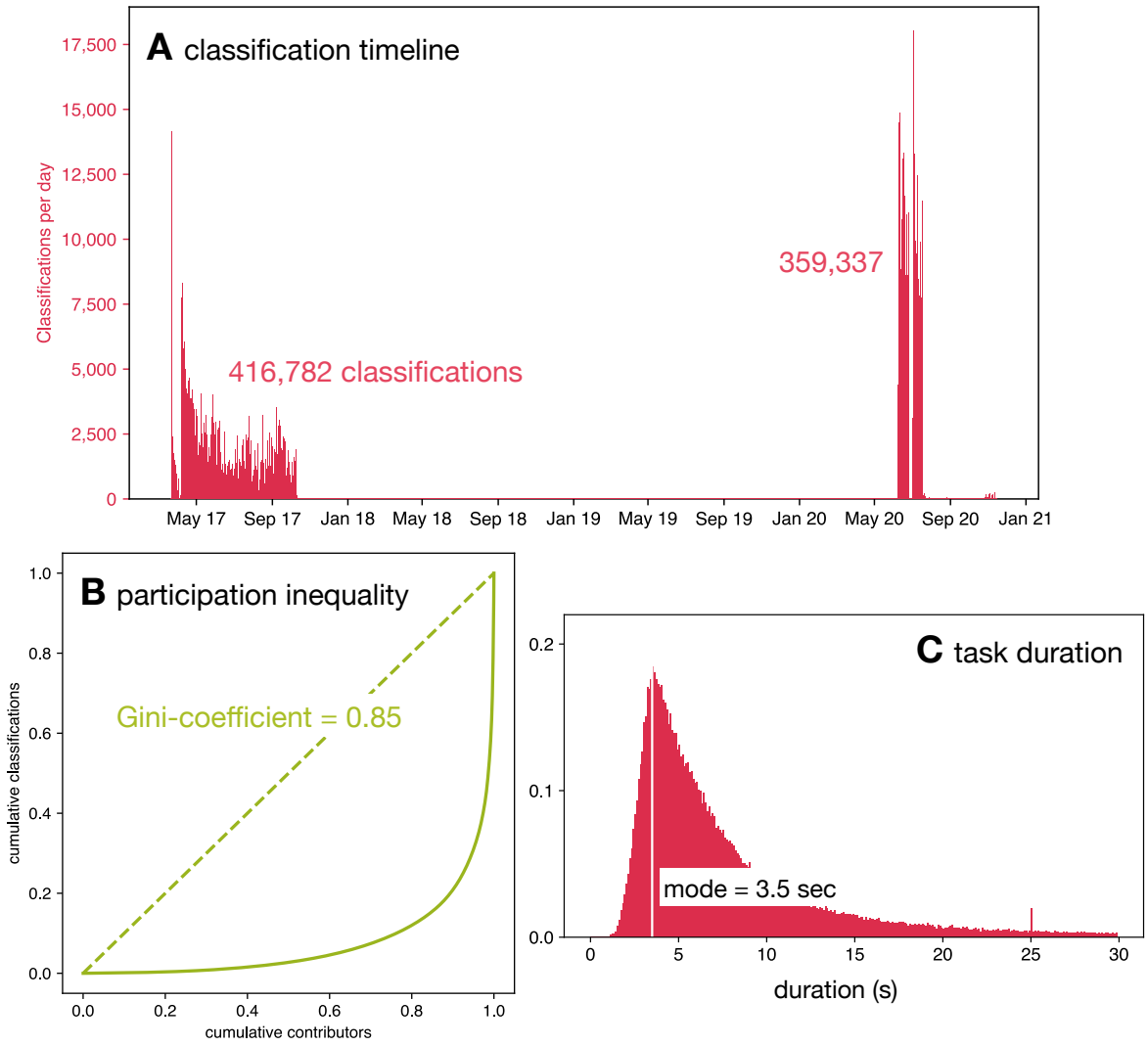

Figure 2: This dataset of 776,119 classifications was collected in two batches between April 2017 and Sep 2020 by 9,062 volunteers. (**A**) The classifications were done by the volunteers in two distinct batches; one during 2017 and a later one in 2020. Note that the higher participation during 2020 was due to the national restrictions imposed due to the SARS-Cov-2 pandemic. (**B**) The Lorenz curve demonstrates that there is considerable participation inequality in the project resulting in a Gini-coefficient of 0.85. (**C**) Volunteers spent different lengths of time classifying drug images after 14 days of incubation with a mode duration of 3.5 seconds.

Time spent

The time spent by a volunteer classifying a single drug image varied from a few seconds up to hours; the latter are assumed to be data artefacts caused by e.g. a volunteer leaving a browser window open. The distribution of time spent per image split shows no appreciable differences when calculated as function of the incubation time with a mode of 3.5 seconds (Fig. 2C, S3), which is unexpected given after only 7 days of incubation there is little or no bacterial growth. After 14 days incubation there are, however, observable differences between how long the volunteers spent classifying each drug (Fig. S5).

Classification validity

The tutorial on the Zooniverse website (Fig. S6) encouraged volunteers to check that the control wells both contain bacterial growth – if not then the drug image should be marked as having "No Growth in either of the 'No Antibiotic' wells". They were also asked to check if any of the drug wells contain growth very different to all the others (contamination), inconsistent growth (skip wells), or anything else that would prevent a measurement being taken (artefacts). If any of these are true, they were asked to mark the drug image as "Cannot classify". In the analysis these were aggregated into a single dilution (NR – not read). In all cases, if a simple majority make a classification of NR, then this is always returned as the result. All NR results are excluded from calculations of the exact or essential agreement.

Expert measurements

Each drug image was also measured by a laboratory scientist using a Vizion instrument as well as programmatically by some software, AMyGDA[10]. Although the AMyGDA software is reproducible, it will often classify artefacts, such as air bubbles, contamination, shadows and sediment, as bacterial growth and is also likely to assess a well as containing no bacterial growth when the level of growth is very low. By contrast, laboratory scientists are not consistent but can recognise and ignore artefacts. Since the sources of error for each these methods are different, we constructed a consensus dataset with an assumed reduced error rate by only including readings where both of these independent methods agree on the value of the MIC. We will refer to this as the 'Expert+AMyGDA' dataset and the larger dataset simply consisting of all the readings taken by the laboratory scientist as the 'Expert' dataset. We shall further assume the error-rate in the 'Expert+AMyGDA' consensus dataset is negligible, allowing us to use it as a reference dataset, which in turn will allow us to infer the performance of the volunteers by comparison.

A total of 12,488 drug images were read after 14 days incubation (Table S3); for 6,205 (49.7%) of these both the laboratory scientist (Expert) and the software (AMyGDA) returned the same MIC. Since a laboratory scientist would be reasonably expected to make an error $\leq 5\%$ of the time, the majority of the drug images excluded are most likely due to AMyGDA incorrectly reading a drug image with only a minority being genuine errors.

By constructing the consensus Expert+AMyGDA dataset we are likely to have introduced bias by unwittingly

selecting for samples which are easier to read. One candidate is that we may have selected samples with higher than average levels of bacterial growth. We can show that this is not the case since not only is the average level of growth in the positive controls (as measured by AMyGDA) for the Expert+AMyGDA dataset (30.8%) similar to that observed (30.6%) for the larger Expert dataset (Fig. S7), but the distributions themselves are very similar.

A second possibility is that drug images with specific growth configurations (for example either no growth or growth in all the wells) are easier to read than drug images where the growth halts. This would imply that the probability of the Expert+AMyGDA measurements agreeing is a function of the dilution MIC, which indeed is what we find (Table S4). The agreement is highest when there is no growth in any of the drug wells, which makes sense as that is a relatively trivial classification to make. The next highest value is when the dilution is 8, which since 7 of the 14 drugs on the plate have 7 drug wells (Fig. 1), corresponds to growth in all 7 drug wells, which is also an easy classification.

The net effect of this is that the Expert+AMyGDA dataset has a different distribution of measured MICs, including a greater proportion of drug images with a low MIC dilution (61.4% after 14 days incubation have a dilution of 1 or 2, Fig. S8) compared to the parent Expert dataset (45.8%). One should bear in mind this bias when interpreting the results, and to assist we will consider if key results change when we use the Expert-only dataset.

How to compare?

Ideally one would apply an international standard for antibiotic susceptibility testing (AST) for Mycobacteria which would permit us to assess if a consensus measurement obtained from a crowd of volunteers is sufficiently reproducible and accurate to be accredited as an AST device. Unfortunately, there is no international standard for Mycobacterial AST – the need to subject Mycobacteria to the same processes and standards as other bacteria has been argued elsewhere[19] – we shall therefore tentatively apply the international standard for aerobic bacteria[20] which requires the results of the proposed antibiotic susceptibility testing method to an appropriate reference method.

Neither of the measurement methods used in constructing our reference consensus dataset has been endorsed, although broth microdilution using Middlebrook 7H9 media was recently selected by EUCAST as a reference method for determining *M. tuberculosis* MICs[21] but only for manually-prepared 96-well plates, permitting much larger numbers of wells for each drug. Nor has any software-based approach for reading MICs from 96-well microdilution plates been endorsed by EUCAST, the CLSI or any other international body. Despite this, and in lieu of any other reasonable approach, we shall treat the consensus MICs (the Expert+AMyGDA datset) as a reference dataset and apply ISO 20776-2[20].

This requires a new AST method that measures MICs to be compared to the reference method using two key metrics: the *exact agreement* and the *essential agreement*. The former is simply the proportion of definite readings which agree, whilst the latter allows for some variability and is defined as the "MIC result obtained by the AST device that is within plus or minus doubling dilution step from the MIC value established by the reference

method"[20]. To meet the standard any new AST method must be $\geq 95\%$ reproducible and $\geq 90\%$ accurate (both assessed using essential agreement) compared to the reference method[20].

Variability in classifications

Inevitably there is a large degree of variation in the classifications made by different volunteers of the same drug image. Examining all 112,163 classifications made by the volunteers on the 6,205 drug images taken after 14 days incubation and comparing them to the consensus of the laboratory scientist and AMyGDA shows that a single volunteer is likely to exactly agree with the Expert+AMyGDA dataset 74.6% of the time, excluding cases where either concluded the drug image could not be read (Table S3). This rises to 86.4% when only considering essential agreement.

The magnitude of agreement varies depending on the measured dilution: if the consensus view is that a drug image contains no growth, a single volunteer is likely to agree 64.1% of the time (Fig. 3A), however this falls to 47.8% if the consensus dataset indicates that the first four wells contain growth before rising to 94.5% when the laboratory scientist decides the dilution is 8. We hence recapitulate our earlier observation that drug images with no growth or growth in all wells are easier to read than drug images where only a subset of drug wells contain growth.

The BashTheBug volunteers are likely to return a higher dilution than the Expert+AMyGDA consensus; this can be seen in the greater proportion of MICs with higher dilutions (Fig. 3B-D). For example a single volunteer is at least $5\times$, and often $> 10\times$, more likely to return an MIC one greater than the reference rather than an MIC one lower than the reference (Fig. 3D). We shall return to this bias later.

When the classification made by an individual volunteer does not agree with the consensus this is often (but not always) because they have misclassified the drug image (Fig. 3E-G). Comparing the classifications made by individual volunteers with the larger, but presumably less accurate, Expert dataset we see that an individual volunteer is less likely to agree with a single laboratory scientist with the overall level of exact agreement falling from 74.6% to 65.3% (Table S3, Fig. S9). Regardless of the comparison dataset used, it is clear that to achieve satisfactory levels of reproducibility and accuracy one must clearly ask *several* volunteers to assess each drug image and then form a consensus measurement which can be compared to the reference measurement. How should we form that consensus?

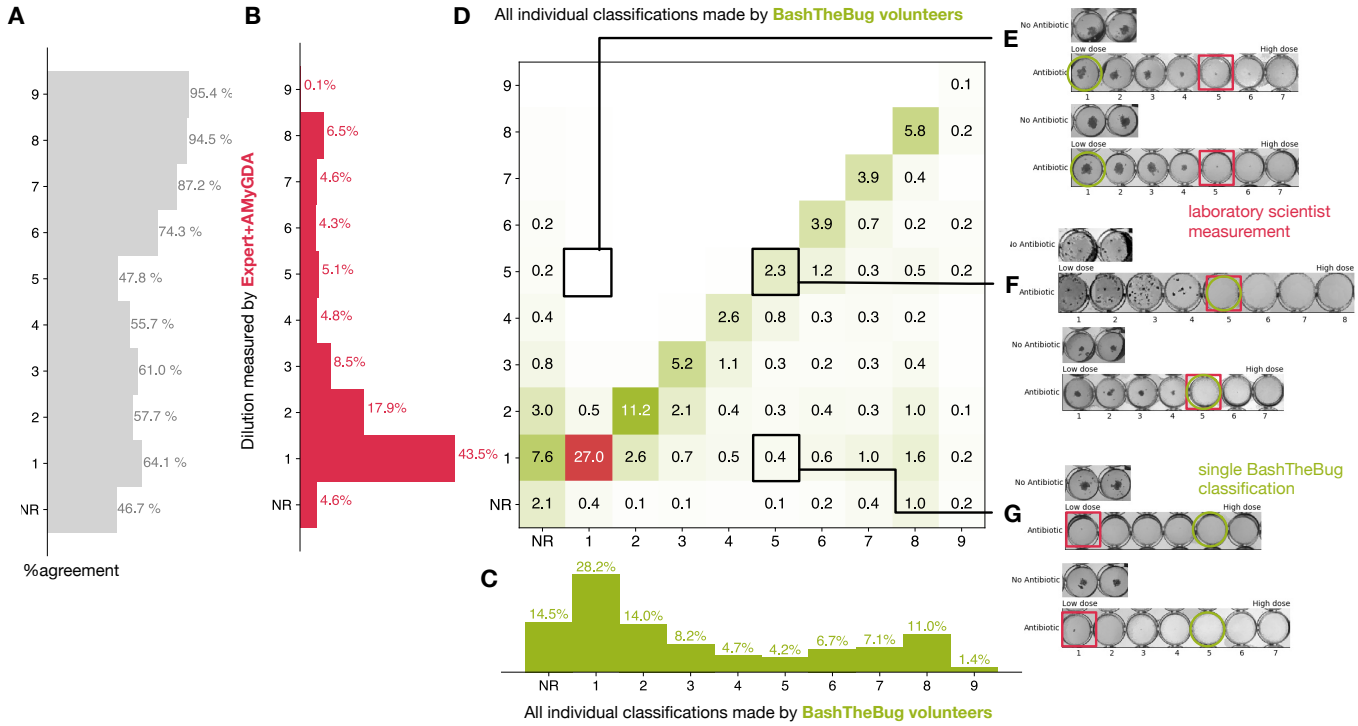

Figure 3: Heatmap showing how all the individual BashTheBug classifications (n=214,164) compare to the dilution measured by the laboratory scientist using the Thermo Fisher Vizion instrument after 14 days incubation (n=12,488) (**A**) The probability that a single volunteer exactly agrees with the Expert+AMyGDA dataset varies with the dilution. (**B**) The distribution of all dilutions in the Expert+AMyGDA dataset after 14 days incubation. The differences are due to different drugs having different numbers of wells as well as the varying levels of resistance in the supplied strains. NR includes both plates that could not be read due to issues with the control wells and problems with individual drugs such as skip wells. (**C**) The distribution of all dilutions measured by the BashTheBug volunteers. (**D**) A heatmap showing the concordance between the Expert+AMyGDA dataset and the classifications made by individual BashTheBug volunteers. Only cells with $> 0.1\%$ are labelled. (**E**) Two example drug images where both the Expert and AMyGDA assessed the MIC as being a dilution of 5 whilst a single volunteer decided no growth could be seen in the image. (**F**) Two example drug images where both the laboratory scientist and a volunteer agreed that the MIC was a dilution of 5. (**G**) Two example drug images where the laboratory scientist decided there was no growth in any of the wells, whilst a single volunteer decided there was growth in the first four wells.

Consensus

There are a range of methods one can use to extract a consensus from a set of classifications; the simplest being majority voting, however, this is not practical since an outright majority is not guaranteed. Alternatively one may take the mode, mean or median of the classifications, although the the former is not always defined and the last two do not always yield an integer. More sophisticated methods, such as the weighted-majority algorithm[22], give weights to the classifiers based on their accuracy, however this requires each volunteer to first classify a ground-truth dataset, which was not available at the start of the project. Given the high level of inequality in participation (Fig. 2D), such methods would be very difficult to apply in practice in our case. We shall therefore limit ourselves here to considering only the mean, median and mode. Since these methods all require the classifications to be numerical, we excluded all readings where the Expert+AMyGDA measurement and/or half or over of the volunteers decided the drug image could not be read. If the classification distribution was bi-modal, then the lower value of the dilution is returned. If necessary, the mean or median were also rounded down.

Reproducibility

To create two consensus measurements by the volunteers of each drug image, two separate sets of 17 classifications were drawn with replacement. By applying the relevant method (mean, median or mode) a consensus dilution was arrived at for each set and then the two results compared. To begin with only drug images with 17 or more classifications were considered and this bootstrapping process was repeated ten times for each drug image in the Expert+AMyGDA dataset. Considering only those drug images taken after 14 days incubation (Fig. 4A & Table S5), they are more likely to exactly agree with one another when the mode was applied ($89.2 \pm 0.1\%$) than the median ($86.9 \pm 0.1\%$) or mean ($74.3 \pm 0.1\%$). For the essential agreement we find that the mean now performs best ($97.2 \pm 0.1\%$), followed by the median ($94.2 \pm 0.1\%$) and mode ($94.1 \pm 0.1\%$).

Hence only the mean exceeds the threshold for reproducibility[20] when 17 classifications are used to build a consensus. Repeating the analysis for the drug images in the larger Expert dataset yields the same conclusion (Fig. S10A). The heatmaps (Fig. 4B) show how two consensus measurements arrived at via the mean tend to be similar but not necessarily identical to one another, whilst two consensus measurements derived using the mode are more likely to agree with one another but also are more likely to arrive at very different values. The median sits in between these two extremes.

Accuracy

Comparing the consensus measurements from the volunteers to the set of MICs in the Expert+AMyGDA dataset yields a different picture (Fig. 4C). The mode exactly agrees with the reference $80.9 \pm 0.1\%$ of the time, followed by the median ($78.1 \pm 0.1\%$) and then mean ($68.4 \pm 0.1\%$). The mean, despite performing best for reproducibility, has the lowest level of essential agreement (as well as exact agreement) with the Expert+AMyGDA readings (88.5

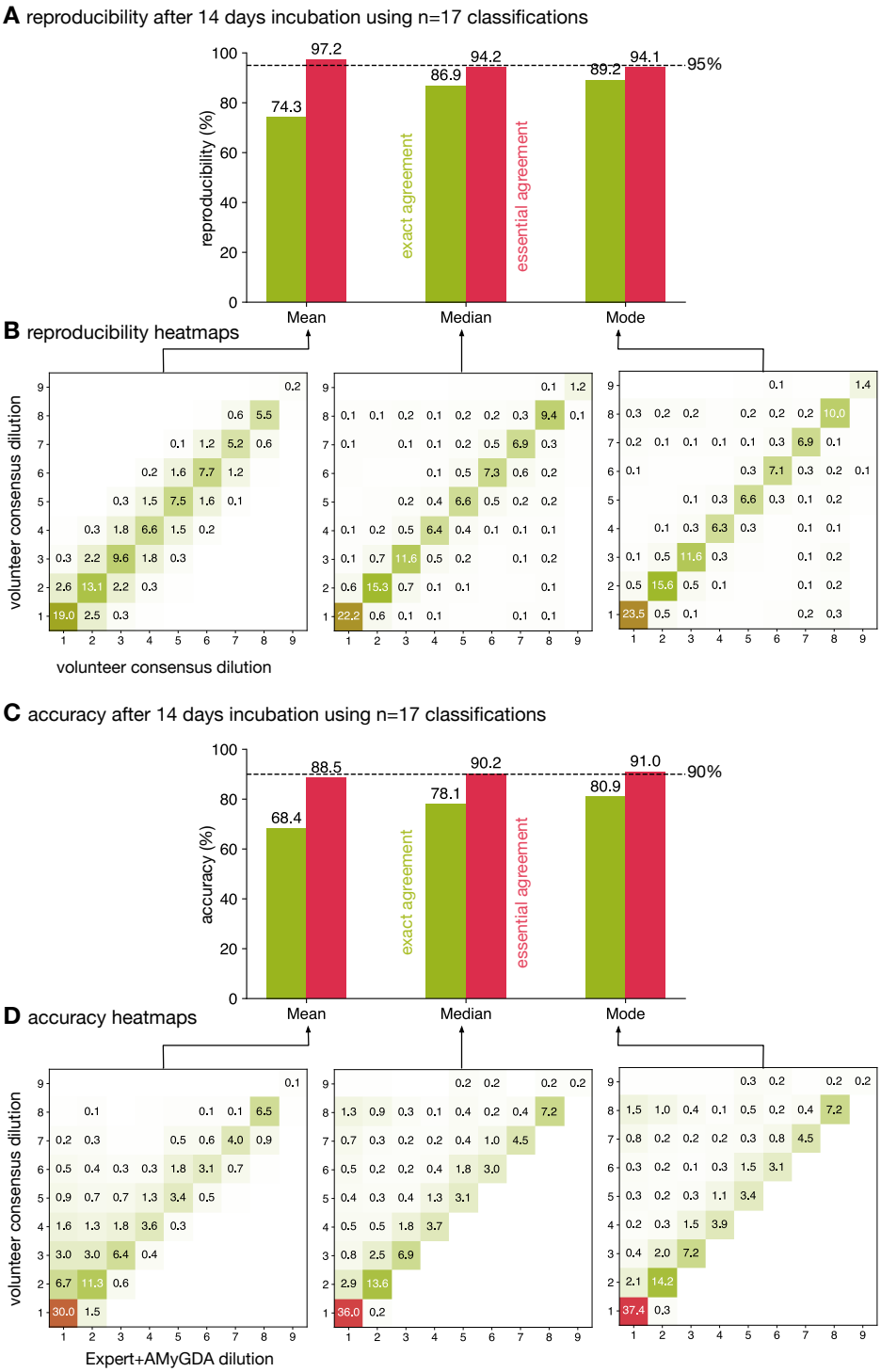

Figure 4: Taking the mean of 17 classifications is $\geq 95\%$ reproducible whilst applying either the median or mode is $\geq 90\%$ accurate. (**A**) Only calculating the mean of 17 classifications achieves an essential agreement $\geq 95\%$ for reproducibility[20], followed by the median and the mode. (**B**) Heatmaps of the consensus formed via the mean, median or mode after 14 days incubation. Only drug images from the Expert+AMyGDA dataset are included. (**C**) The essential agreement between a consensus dilution formed from 17 classifications using the median or mode and the consensus Expert+AMyGDA dilution both exceed the required 90% threshold[20]. (**D**) The heatmaps clearly show how the volunteer consensus dilution is likely to be the same or greater than the Expert+AMyGDA consensus.

$\pm$ 0.1%), with the median (90.2 $\pm$ 0.1%) and mode (91.0 $\pm$ 0.1%) both exceeding the 90% accuracy threshold[20].

The heatmaps show how the consensus dilution of the classifications made by the volunteers is much more likely to be higher than the Expert+AMyGDA measurement than lower (Fig. 4D), regardless of the consensus method, indicating perhaps that volunteers are more likely than laboratory scientists to classify any dark regions in a well as bacterial growth, or that laboratory scientists are more willing to discount some features as artefacts e.g. air bubbles or sediment. Repeating the analysis using the Expert dataset (Fig. S10) leads to lower values for the exact and essential agreements for all consensus methods – this is to be expected since this the Expert reference dataset contains a larger proportion of errors than the Expert+AMyGDA dataset.

Which method to choose?

Despite being the most reproducible method as measured by essential agreement, we discount the mean since it suffers from relatively poor levels of exact agreement for both reproducibility and accuracy and its performance falls faster than the other methods when n is decreased. The median and mode have very similar reproducibilities and accuracy and we conclude they perform equally well. We can infer from this that bi-modal classification distributions are rare and that the median is often identical to the mode.

Reducing the number of classifications

Clearly it would be desirable and ethical to only require the volunteers to complete the minimum number of classifications to achieve an acceptable result. The simplest way to do this is to decrease the number of classifications, n, before a drug image is retired – this reduces both the reproducibility and accuracy of the consensus measurements (Fig. 5, Table S5, S6), however perhaps not by as much as one might expect. The mean exceeds the essential agreement $\geq$ 95% reproducibility threshold for $n \geq 13$, whilst the mode and the median satisfy the accuracy criterion of essential agreement $\geq$ 90% for $n \geq 3$ and $n \geq 11$, respectively (Fig. 5, Table S6). Similar trends are observed when the Expert dataset is used as the reference (Fig. S12). Accuracy is hence less sensitive than reproducibility to reducing the number of classifications used to build the consensus and, depending on the consensus method used, the number of classifications can be reduced whilst still maintaining acceptable levels of accuracy.

Can we improve matters?

Retiring all drug images after a fixed number of classifications is simple but does not take account of the relative difficulty of the classification task. If one was able to group the drug images by difficulty, either before upload or dynamically during classification, then one could optimally target the work undertaken by the volunteers. Due to the inherent difficulties in culturing *M. tuberculosis*, there is a broad distribution of growth in the positive control wells after 14 days incubation (Fig. S7). One might expect that drug images with poor growth would be more challenging to classify, however, segmenting by low, medium and high growth shows the amount of growth in the positive control wells has little effect on either the reproducibility (Fig. S14) or accuracy (Fig. S15), regardless of

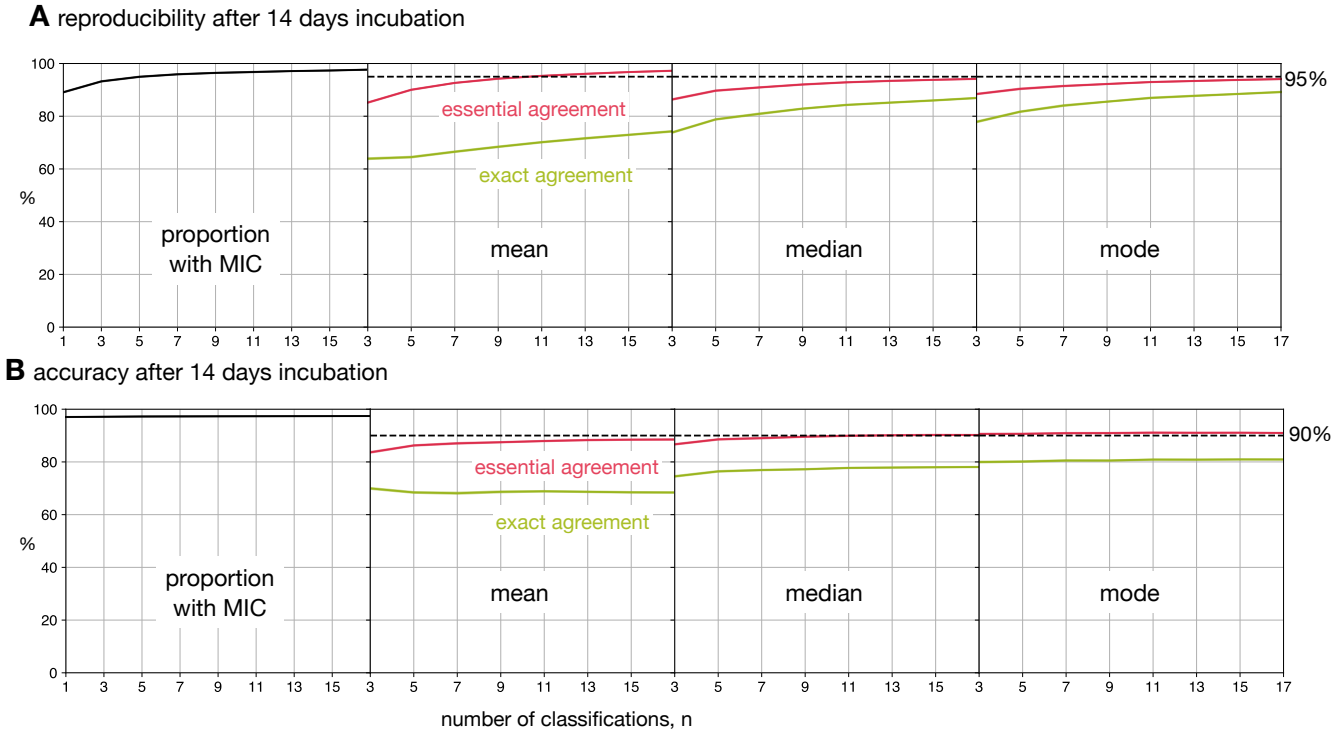

Figure 5: Reducing the number of classifications, n, used to build the consensus dilution decreases the reproducibility and accuracy of the consensus measurement. (**A**) The consensus dilution becomes less reproducible as the number of classifications is reduced, as measured by both the exact and essential agreements. (**B**) Likewise, the consensus dilution becomes less accurate as the number of classifications is decreased, however the highest level of exact agreement using the mean is obtained when $n = 3$ and the mode, and to a lesser extent the median, are relatively insensitive to the number of classifications. These data are all with respect to the Expert+AMyGDA dataset.

the consensus method and number of classifications employed.

Alternatively one could use the first few classifications performed by the volunteers to assess the difficulty of each drug image. For example, if the first n volunteers all return the same classification, then it is reasonable to assume that this is a straightforward image and it can be retired, with the remainder accruing additional classifications. Ideally one would want to develop an algorithm that assessed the likelihood of the classification not being altered by more classifications to allow a dynamic decision about when to halt, however applying such an approach is not yet possible within the Zooniverse.

To estimate the potential value in applying a simple approach to dynamically retiring drug images, we shall consider applying the median after 14 days of incubation and will arbitrarily retire a drug image if the first three volunteers all made the same classification, with all other drug images being retired after 17 classifications. This simple protocol reduces the number of classifications required to $n = 8.8$, a reduction of 48%, and the reproducibility, as measured by exact agreement, rises from 86.8% to 87.6%, whilst the essential agreement remains unchanged (94.2% to 94.4%). The accuracy, assessed in the same way, behaves similarly with the exact agreement increasing from 78.1% to 78.8% with the essential agreement remaining unaltered (90.2% to 90.3%). Hence retiring some of the drug images at $n = 3$ not only dramatically reduces the number of classifications required but also improves the result in a few cases, presumably because the subsequent classifications have a small chance of altering the dilution by a single unit, hence worsening the exact agreement but not affecting the essential agreement.

A fairer test is to ask if this dynamic approach improves performance if we are constrained to a fixed number of total classifications: if we choose $n = 9$, then the reproducibility of the median (as measured by exact and essential agreements) improves from 83.0% & 92.2% to 87.6% & 94.4% and the accuracy, measured in the same way, improves slightly from 77.2% & 89.6% to 78.8% & 90.3%. We therefore conclude that even a simple dynamic approach to retiring images would minimise the work done by the volunteers / allow more images to be classified.

Variation by drug

So far we have analysed the reproducibility and accuracy of consensus MICs obtained from a crowd of volunteers, thereby aggregating the results for each of the 13 anti-tuberculars (excl. PAS) present on the UKMYC5 plate design. The reproducibility of each drug, as measured by the exact and essential agreements, varies between 79.6-91.4% and 90.8-97.0%, respectively (Fig. 6A). Previous analysis showed that the reproducibility of the whole plate under these conditions when assessed using the essential agreement is 94.2 ± 0.1% (Fig. 4A) – this is below the 95% threshold specified by an international standard for aerobic bacteria[20]. Applying the same threshold to each drug we find that five out of the 13 drugs meet or exceed the threshold whilst the plate as a whole does not. The accuracy of each drug varies more widely: between 48.4-88.9% and 82.8-94.7% when assessed using the exact and essential agreement, respectively. Hence whilst the accuracy of the plate as a whole was 90.2 ± 0.1%, just exceeding the 90% threshold, only six out of 13 drugs surpassed the same threshold.

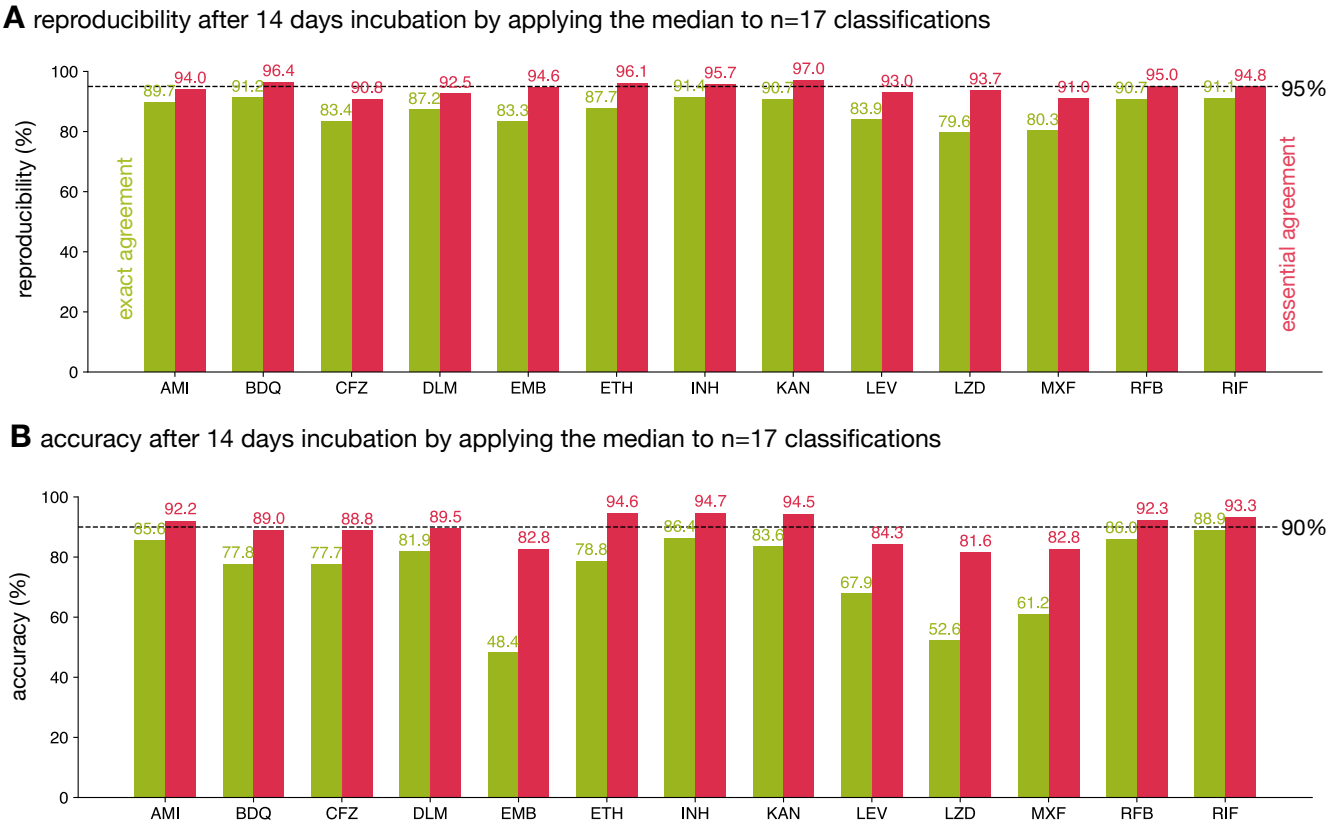

Figure 6: The reproducibility and accuracy of the consensus MICs varies by drug. Consensus MICs were arrived at by taking the median of 17 classifications after 14 days incubation. The essential and exact agreements are drawn as red and green bars, respectively. For the former the minimum thresholds required are 95% and 90% for the reproducibility and accuracy, respectively[20]. See Fig. S16 for the other consensus methods.

The variation in reproducibility and accuracy between anti-tuberculars, as well as between the exact and essential agreement for a single compound, is due to a number of factors, not limited to the number and concentration range of the wells on the plate design, the mechanisms of both action and resistance, the prevalence of resistance and the degradation rate after lyophilisation, both during storage and after inoculation. For example, kanamycin passes the reproducibility and accuracy thresholds we have adopted and this is likely due to there being relatively few (five) drug wells on the UKMYC5 plate (Fig. S2) and the mechanism of resistance being such a substantial proportion of samples either do not grow in any well, or grow in all the drug wells, making measurement more straightforward. The mechanism of action of each compound is likely to affect how 'easy' it is to determine the minimum inhibitory concentration. From the striking differences in exact and essential accuracies of reading ethambutol, moxifloxacin and linezolid we hypothesise the Mycobacterial growth diminishes more gradually with increasing concentration for these drugs, rather than coming to an abrupt end, as it does for other compounds.

This whole analysis could be considered somewhat moot since the UKMYC5 96-well plate would be treated as a single entity (or medical device) if accreditation were to be sought and therefore the results for individual compounds would likely not be considered. One unintended consequence of the current standards is therefore that one could improve the performance of a plate design by dropping compounds with lower-than-average performance, even if this is clinically not desirable, rather than work to e.g. improve the performance of the measurement methods.

DISCUSSION

A crowd of volunteers can reproducibly and accurately measure the growth of a bacterial pathogen on a 96-well broth microdilution plate, thereby demonstrating the potential for clinical microbiology to embrace and combine contrasting measurement methods. No Mycobacterial antibiotic susceptibility testing standard exists, although efforts are underway to establish a reference method[21], and so we applied the standard for antibiotic susceptibility testing of aerobic bacteria[20]. Forming a consensus by applying the mode or median to 17 independent classifications performs better overall than the mean, and the reproducibility of both these methods, as measured by the essential agreement, is 94.2% and 94.1%, respectively (Fig. 4). This is slightly less than the 95% threshold set by ISO for aerobic bacteria and therefore the volunteers do not need this criterion. The accuracy of the crowd, as measured by the essential agreement, is 90.2% and 91.0% when the median and mode, respectively, are applied to produce a consensus measurement – these values are above the required 90% threshold[20], and therefore the volunteers are sufficiently accurate (but not quite reproducible enough) to be classified as an antibiotic susceptibility testing (AST) device.

The volunteers are fast, taking on average 3.5 seconds per drug image, and therefore a single plate requires slightly less than 13 minutes of volunteer time to read if 17 classifications are amassed for each drug. Reducing the number of classifications before an image is retired reduces the reproducibility and accuracy, but not by as much as

one might expect. A more nuanced approach would be to retire a drug image early if the first few classifications are identical, however it is not yet possible to define this type of dynamical rule in the Zooniverse portal. The level of participation by the volunteers was very unequal with a small cadre of volunteers doing very large numbers with ten volunteers doing, on average, over 10,000 classifications each which is more than many of the laboratory scientists who are considered the experts! Compared to the measurements taken by the laboratory scientists, the consensus dilution arrived at by the volunteers tends to be higher, indicating a bias to overcall (Fig. S10D), which is supported by anecdotal observations of people classifying drug images at public engagement events where they often choose a higher dilution 'to be on the safe side'. By contrast, the AMyGDA computer software has been noted to have the opposite bias – i.e. be more likely to undercall compared to the expert[10]. These oppositely directly biases will make it more difficult to use all three methods to reduce the level of measurement error in large datasets since they reduce the likelihood that different measurements methods will exactly agree with one another.

The reproducibility and accuracy of any method used to read a 96-well microtitre plate, whether that is laboratory scientists using a Thermo Fisher Vizion instrument or citizen scientists visually examining drug images, depends on a range of factors from the prevalence of drug-resistant samples in the dataset to which drugs are included in the plate design and the number and concentrations of their allotted wells. For the UKMYC5 plate design, both the Expert and either the AMyGDA or BashTheBug measurements are more likely to agree with one another at low dilutions where there is little or no *M. tuberculosis* growth in the drug wells (Tables S4, S9), hence reducing the number of resistant samples would artificially 'improve' performance yet the standards do not specify the degree of resistance in any test dataset[20]. The requirement to have quality control strains that have definite growth in all the drug wells unintentionally mitigates against this risk. The dataset used here was based on 19 external quality assessment strains and therefore whilst it included some degree of resistance for all 13 drugs, there was only a single strain resistant to clofazimine, bedaquiline or delamanid and no strain was resistant to linezolid[9]. For isoniazid and rifampicin, 8 and 7 of the 19 EQA strains, respectively were resistant and hence the prevalence of resistance for these drugs is much greater than would be expected to be encountered in most countries. Clearly studies including a much more diverse range of strains, for example clinical isolates, would be more definitive. Since the 13 antituberculars on the UKMYC5 plate (Fig. S2) also all perform differently (Fig. 6) different plate designs will perform differently, which is important as it is the plate that would be accredited, rather than the individual compounds.

Although the primary aim of this study was to assess whether the measurements produced by a crowd of volunteers are sufficiently reproducible and accurate to help reduce the measurement error in datasets containing large numbers of microtitre plates (as is being collected by the CRyPTIC project and others) the resulting dataset of classifications is ideally suited to train machine-learning models. This is increasingly recognised as an important use of citizen science[23] and one could envisage training a light-weight machine-learning algorithm able to run on a

mobile device which, by taking a photograph of a 96-well plate, could automatically read the minimum inhibitory concentrations. The best use of such a device would likely be to act as a double check for readings taken by the laboratory scientist. Alternatively, one could build a hybrid approach where e.g. small crowds of experts could examine plates used in a clinical microbiology service – these could be particularly difficult drug images or could be a random sample for quality assurance purposes. This type of hybrid approach would also help with training laboratory scientists which would help reduce the barrier to using 96-well microtitre plates for *M. tuberculosis* AST in clinical microbiology laboratories, especially in low- and middle-income countries. Finally, it is likely that each volunteer has their own individual bias and variability and constructing consensus methods[24] that take these into account would likely further improve the performance of crowds of citizen scientists.

FUNDING

BashTheBug was supported by Welcome through the Enriching Engagement Grants scheme at the University of Oxford. This work was also supported by Wellcome Trust/Newton Fund-MRC Collaborative Award [200205/Z/15/Z]; and Bill & Melinda Gates Foundation Trust [OPP1133541]. This study is funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, a partnership between Public Health England and the University of Oxford. The research was also funded/supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including from the National Science Foundation, NASA, the Institute of Museum and Library Services, UKRI, a Global Impact Award from Google, and the Alfred P. Sloan Foundation.

ACKNOWLEDGEMENTS

We are very grateful to all the volunteers (Fig. S1) who contributed their time and energy to this project and to the Zooniverse development team for coding and maintaining the Zooniverse online platform. We thank David Hawkins for designing the BashTheBug logo and typeface and Chris Wood, Oxford Medical Illustration and Dr Nicola Fawcett, livinginamicrobialworld.com for the wild garden of the gut bacteria photographs. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. World Health Organization (2020) Global Tuberculosis Report. Technical report.

2. O'Neill J (2016) Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. Technical report.

3. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ, Clifton DA, Kapatai G, Ip CLC, Bowden R, Drobniewski FA, Allix-Béguec C, Gaudin C, Parkhill J, Diel R, Supply P, Crook DW, Smith EG, Walker AS, Ismail N, Niemann S, Peto TEA, Modernizing Medical Microbiology (MMM) Informatics Group (2015) *Lancet Infec Disease* 15:1193–202.

4. Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks AM, Emerson C, Hanna D, Kim PS, Liwski R, Zignol M, Gilpin C, Niemann S, Denkinger CM, Fleming J, Warren RM, Crook D, Posey J, Gagneux S, Hoffner S, Rodrigues C, Comas I, Engelthaler DM, Murray M, Alland D, Rigouts L, Lange C, Dheda K, Hasan R, Ranganathan UDK, McNerney R, Ezewudo M, Cirillo DM, Schito M, Köser CU, Rodwell TC (2017) *Eur Respir J* 50:1701354.

5. The CRyPTIC Consortium, 100000 Genomes Project (2018) *New Eng J Med* 379:1403–1415.

6. Pankhurst LJ, del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, Fermont JM, Gascoyne-Binzi DM, Kohl TA, Kong C, Lemaitre N, Niemann S, Paul J, Rogers TR, Roycroft E, Smith EG, Supply P, Tang P, Wilcox MH, Wordsworth S, Wyllie D, Xu L, Crook DW (2016) *Lancet Resp Med* 4:49–58.

7. Walker TM, Cruz ALG, Peto TE, Smith EG, Esmail H, Crook DW (2017) *Lancet Infec Disease* 17:359–361.

8. Kubica GP, Kim TH, Dunbar FP (1972) *Int J System Bacteriol* 22:99–106.

9. Rancoita PMV, Cugnata F, Gibertoni Cruz AL, Borroni E, Hoosdally SJ, Walker TM, Grazian C, Davies TJ, Peto TEA, Crook DW, Fowler PW, Cirillo DM, Crook DW, Peto TEA, Walker AS, Hoosdally SJ, Gibertoni Cruz AL, Grazian C, Walker TM, Fowler PW, Wilson D, Clifton D, Iqbal Z, Hunt M, Smith EG, Rathod P, Jarrett L, Matias D, Cirillo DM, Borroni E, Battaglia S, Chiacchiaretta M, De Filippo M, Cabibbe A, Tahseen S, Mistry N, Nilgiriwala K, Chitalia V, Ganesan N, Papewar A, Rodrigues C, Kambli P, Surve U, Khot R, Niemann S, Kohl T, Merker M, Hoffmann H, Lehmann S, Plesnik S, Ismail N, Omar SV, Joseph L, Marubini E, Thwaites G, Thuy Thuong TN, Ngoc NH, Srinivasan V, Moore D, Coronel J, Solano W, He G, Zhu B, Zhou Y, Ma A, Yu P, Schito M, Claxton P, Laurenson I (2018) *Antimicrobial Agents and Chemotherapy* 62:e00344–18.

10. Fowler PW, Gibertoni Cruz AL, Hoosdally SJ, Jarrett L, Borroni E, Chiacchiaretta M, Rathod P, Lehmann S, Molodtsov N, Grazian C, Walker TM, Robinson E, Hoffmann H, Peto TEA, Cirillo DM, Smith GE, Crook DW (2018) *Microbiology* 164:1522–1530.

11. McMaster A, Hutchings R, Allen C, Wolfenbarger Z, Dickinson H, Trouille L, Johnson C (2021). Panoptes CLI. A command-line interface for Panoptes, the API behind the Zooniverse.

12. Fowler PW (2018). pyniverse: a Python package to analyse classifications made by volunteers in a generic Zooniverse citizen science project. https://github.com/philipwfowler/pyniverse.

13. Fowler PW (2018). bashthebug: a Python package to analyse the results of the Zooniverse volunteers for the BashTheBug citizen science project. https://github.com/philipwfowler/bashthebug.

14. McKinney W (2010) In Proceedings of the 9th Python in Science Conference, edited by SvdW Millman, Jarrod, 51–56.

15. Fowler PW (2017). https://bashthebug.net/ - Help us fight antibiotic resistance!

16. Spiers H, Swanson A, Fortson L, Simmons BD, Trouille L, Blickhan S, Lintott C (2019) *Journal of Science Communication* 18:1–32.

17. Cox J, Oh EY, Simmons B, Lintott C, Masters K, Greenhill A, Graham G, Holmes K (2015) *Computing in Science and Engineering* 17:28–41.

18. Spiers H, Songhurst H, Nightingale L, de Folter J, Hutchings R, Peddie CJ, Weston A, Strange A, Hindmarsh S, Lintott C, Collinson LM, Jones ML (2020) *bioRxiv* 1–41.

19. Schön T, Matuschek E, Mohamed S, Utukuri M, Heysell S, Alffenaar JW, Shin S, Martinez E, Sintchenko V, Maurer F, Keller P, Kahlmeter G, Köser C (2019) *Clinical Microbiology and Infection* 25:403–405.

20. International Organization for Standardization (2007) ISO 20776-2: Clinical laboratory testing and in vitro diagnostic test systems. Technical report, International Standards Organization.

21. Schön T, Werngren J, Machado D, Borroni E, Wijkander M, Lina G, Mouton J, Matuschek E, Kahlmeter G, Giske C, Santin M, Cirillo DM, Viveiros M, Cambau E (2020) *Clinical Microbiology and Infection* 27:10–13.

22. Littlestone N, Warmuth M (1989) In 30th Annual Symposium on Foundations of Computer Science, 256–261. IEEE.

23. Trouille L, Lintott CJ, Fortson LF (2019) *Proc Natl Acad Sci U S A* 116:1902–1909.

24. Zhu T, Johnson AE, Behar J, Clifford GD (2014) *Annals of Biomedical Engineering* 42:871–884.