The first AI-based mobile application for antibiotic resistance testing

Marco Pascucci[1,2,9], Guilhem Royer[3,4,9], Jakub Adamek[6], David Aristizabal[6], Laetitia Blanche[1], Amine Bezzarga[1,8], Dominique Boissinot[7], Guillaume Boniface-Chang[6], Alex Brunner[6], Christian Curel[7], Gabriel Dulac-Arnold[5], Nada Malou[1], Clara Nordon[1], Vincent Runge[2], Franck Samson[2], Ellen Sebastian[6], Dena Soukieh[6], Jean-Philippe Vert[5], Christophe Ambroise[2], Mohammed-Amin Madoui[3]

[1]*Fondation Médecins Sans Frontières, Paris, France*
[2]*Université d'Évry, Université Paris-Saclay, LaMME, Évry, France*
[3]*Génomique Métabolique, CEA, Genoscope, Institut François Jacob, CNRS, Université d'Évry, Université Paris-Saclay, Évry, France*
[4]*Département Prevention, Diagnostique, Traitement des Infections, Hôpital Universitaire Henri Mondor, Assistance Publique- Hôpitaux de Paris, Créteil, France*
[5]*Google Research, Brain Team, Paris, France*
[6]*Google, Mountain View, CA, United-States*
[7]*i2a, Montpellier, France*
[8]*X-Squad, Paris, France*
[9]*authors with equal contributions*

Antimicrobial resistance is a major global health threat and its development is promoted by antibiotic misuse. While disk diffusion antibiotic susceptibility testing (AST, also called antibiogram) is broadly used to test for antibiotic resistance in bacterial infections, it faces strong criticism because of inter-operator variability and the complexity of interpretative reading. Automatic reading systems address these issues, but are not always adapted or available to resource-limited settings. We present the first artificial intelligence (AI)-based, offline smartphone application for antibiogram analysis. The application captures images with the phone's camera, and the user is guided throughout the analysis on the same device by a user-friendly graphical interface. An embedded expert system validates the coherence of the antibiogram data and provides interpreted results. The fully automatic measurement procedure of our application's reading system achieves an overall agreement of $90\,\%$ on susceptibility categorization against a hospital-standard automatic system and $98\,\%$ against manual measurement (gold standard), with reduced inter-operator variability. The application's performance showed that the automatic reading of antibiotic resistance testing is entirely feasible on a smartphone. Moreover our application is suited for resource-limited settings, and therefore has the potential to significantly increase patients' access to AST worldwide.

Introduction

The development of new antimicrobial agents is currently outpaced by the emergence of new antimicrobial resistance[1] (AMR). The development and diffusion of AMR has become a serious health threat[2], whose magnitude is not yet fully understood because of the lack of data, especially in areas where the access to antimicrobial susceptibility testing is difficult. A high-profile review[3] forecasts ten million deaths worldwide by 2050. Although these numbers have been criticized[2], these studies underline the critical health burden of AMR and the need for global data[2,4].

Testing the susceptibility of bacteria is important for patient treatment and, if done systematically, gathering data can provide precious epidemiological information. Different test methods[5] exist. Arguably the most widely used is the Kirby-Bauer disk diffusion test.

In this test, cellulose disks (pellets) containing antibiotics at a given concentration are placed in a Petri dish with an agar-based growth medium previously inoculated with bacteria. While the plate is left to incubate, the antibiotic diffuses from the pellet into the agar. The antibiotic concentration is highest near a pellet and decreases radially as the distance from the disk increases[6]. The bacteria cannot grow around those disks that contain antibiotics to which they are susceptible. The growth of the bacterial colony stops at a distance from the pellet which corresponds to a critical antibiotic concentration, forming a visible bacteria-free area around the cellulose disk. This is called a *zone of inhibition*. After incubation, the diameter of the zone of inhibition around each antibiotic disk is measured: the categorization of the microorganisms as susceptible (S) Intermediate (I) or Resistant (R) is obtained by comparison of the diameter against standard breakpoints[7] established by international committees such as the European Committee on Anitimicrobial Susceptibility Testing (EUCAST) or the Clinical and Laboratory Standards Institute (CLSI) [8].

The disk diffusion method is relatively simple, can be performed entirely by hand, requires no advanced hardware, and has a low cost. However, it is criticized for several reasons. First, it is labor-intensive and time-consuming. Second, it is subject to important inter-operator variability: accurate performance of disk diffusion testing relies on proficient technicians, starting with the quality of plate preparation (e.g. inoculum, purity)[9]. The diameter of the inhibition zone is measured by eye with a caliper or ruler and approximated to the closest mm[10]. However, the inhibition zone might not be a perfect disk, e.g. if the inhibition zone overlap, or if the pellet is too close to the border of the dish (Figure 4a, f and h). In this case the problem of measuring a diameter is ill-posed and, together with intrinsic measurement error, introduces subjectivity and inter-operator variability in the measurement. Third, it requires an advanced level of expertise for interpretation. Sometimes, the inhibition zone diameter is not sufficient by itself to determine the susceptibility. Indeed, several mechanisms of resistance are expressed at a low level *in vitro* but have major impact *in vivo* and can lead to treatment failure. Moreover, susceptibility to a whole class of antibiotics or a given molecule class can sometimes be inferred from the susceptibility to

another one, thus reducing the number of required tests. In those cases, interpretative reading is needed. Interpretation is based on expert rules published and updated by scientific societies, such as EUCAST in Europe[11].

Automatic reading systems have been introduced to alleviate the drawbacks of disk diffusion AST [12,13]. These systems acquire pictures of the plate and automatically measure the diameters of the inhibition zones. Most of them include an *expert system* which can elaborate interpreted results. It helps mitigate the risk that the laboratory reports erroneous susceptibility results and ensures compliance with regulatory guidelines. Commercial devices that automatically read antibiograms are commonly used in hospitals and laboratories[14,15], but the procedures they use are not fully disclosed.

Because of their price, and hardware and infrastructure requirements, these systems are not suited to resource-limited settings such as dispensaries or hospitals in low and middle income countries. Affordable solutions are few. Image processing algorithms for automatic measuring inhibition diameters have been published [16–20]. Among these, only AntibiogramJ[20] presents a fully functional user-friendly software, but it operates on a desktop computer onto which the images need to be previously transferred. We believe that reducing hardware requirements to just a smartphone is key for the adoption and diffusion of such a tool. Moreover, smartphone applications are easy to adopt and use if they follow established design patterns, and they benefit from an ecosystem which facilitates setup and updates.

This paper introduces the first fully offline mobile application (the App hereafter) capable of analyzing disk diffusion ASTs and yielding interpreted results, operating entirely on a smartphone. The need of such an application was identified by Médecins Sans Frontières, who often operates in low and middle income conuntries (LIMC) where AST is difficult or impossible to implement. The MSF Foundation brought together the peoples and skills needed for this application to be developed. This MSF application can have a great impact for the organization and the global fight against AMR. the App combines original algorithms, using machine learning (ML) and image processing, with a rule-based expert system, for automatic AST analysis (see Figure 1). It embeds a clinically tested third-party expert system[13,15] which could compensate for a lack of microbiology expertise. The user is guided throughout the whole analysis and can interact at any step with the user-friendly graphical interface of the application to verify and possibly correct the automatic measurements if needed.

The whole analysis takes place on the same smartphone used to take the picture of the AST. Since it does not require any hardware other than a basic Android smartphone, and because it works completely offline (without internet connection), the App is suited for resource-limited settings. Therefore, the App could help fill the digital gap and increase patients' access to AST worldwide. Finally, if integrated to a global health surveillance program, it could facilitate the collection of

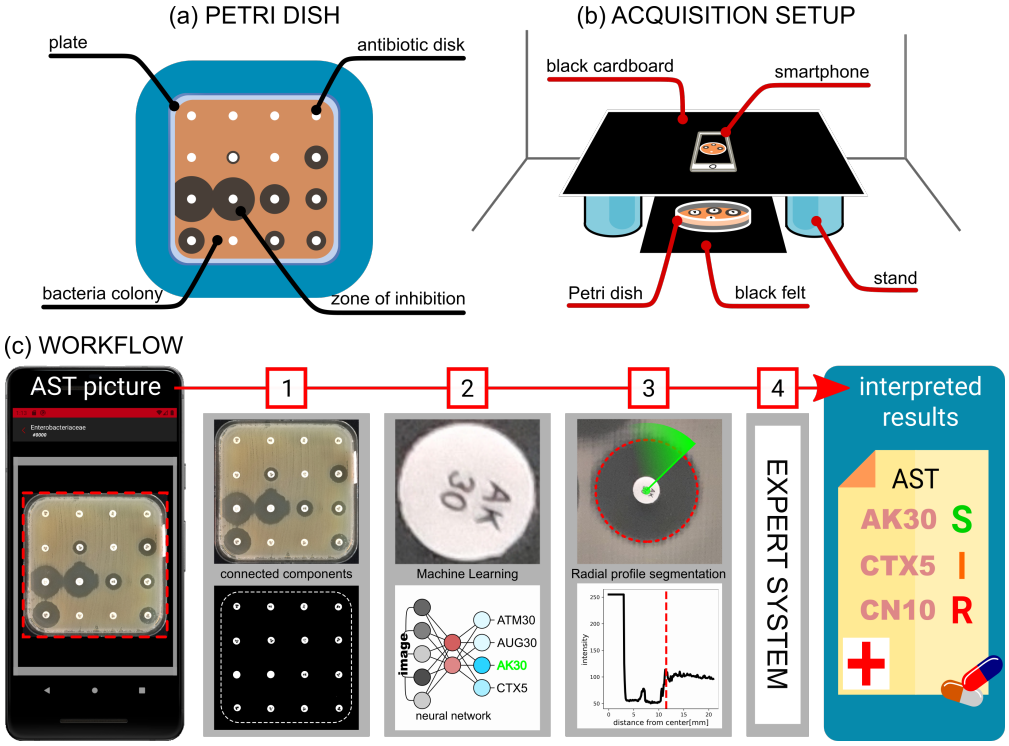

Figure 1: Analysis steps with the App. A prepared an incubated petri dish (a) is positioned in a simple carboard setup (b). Then a picture is taken with a smartphone analysis follows the workflow described in (c). The Petri dish image is cropped and the antibiotic disks are found (c1). The image of each antibiotic disk is fed to an ML model that identifies the antibiotic (c2). Successively, the diameter of the inhibition zone is measured (c3). The embedded expert system uses the diameters to output interpreted results (c4).

epidemiological data on antimicrobial resistances, the lack of which is recognized today as a major health danger[2,4].

In the following, we demonstrate that the App's performance is similar to that obtained with a commercial system and very close to manual reading (considered as the gold standard[10]). The application's full automatic procedure is evaluated on antibiograms prepared in laboratory conditions both on standard and blood-enriched agar. Moreover, we explore the feasibility of a ML-based automatic detection of resistance mechanisms recognizable by peculiar shapes of the inhibition zones.

Results and discussion

the App is the first automatic AST reading system capable of running the whole analysis of a disk-diffusion antibiogram offline on mobile devices, from image acquisition to interpreted results. It helps laboratory technicians throughout the whole analysis process, suggesting measurements, results and interpretations. the App can be summarized in three major components:

- an image processing module (IP) that reads and analyzes the AST image
- an expert system (ES) responsible for the interpretation of the data extracted by IP
- a Graphical User Interface (GUI) that allows the execution of IP and ES on a smartphone.

The application's IP module is a C++ library developed with OpenCV[21] and Tensorflow[22] external libraries. The choice of C++ makes our library exploitable in various contexts, including desktop computers and Android and iOS mobile devices. the App includes an expert system capable of performing coherence checks on the raw susceptibility and providing interpreted results, with extrapolation on non-tested antibiotics and clinical commentaries. The expert system's knowledge base is provided and regularly updated by i2a (Montpellier, France)[13] and based on up-to-date EUCAST expert rules[11]. The expert system's engine was completely developed in TypeScript in order to be operated offline by the App.

Commercial AST reading systems use built-in image acquisition devices (cameras and scanners) to ensure input consistency. the App works on images of ASTs taken directly with the phone's camera, with no additional external acquisition hardware. This inevitably introduces a certain variability in the image quality. We tackle this issue by introducing a simple set of guidelines for image acquisition. These guidelines are designed to optimize image quality, and therefore to reduce the need of heavy post-processing and the risk of numerical artifacts. Moreover, they are conceived to be inexpensive and easy to implement and integrate in the laboratory routine. Since smartphone cameras are not designed for quantitative measurements, we provide a simple method to assess

AST set	Control	Images	Antibiotics	% Agreement	disagreement (%)			kappa
					very major	major	minor	
A1								
overall	SIRscan	570	7334	90	54 (0.7)	428 (5.8)	270 (3.7)	0.77
standard	SIRscan	561	7223	90	46 (0.6)	390 (5.4)	269 (3.7)	0.77
problematic	SIRscan	9	111	58	8 (7.2)	38 (34.2)	1 (0.9)	0.12
A2								
overall	SIRscan	75	534	91	4 (0.7)	36 (6.7)	6 (1.1)	0.71
standard	SIRscan	73	509	95	4 (0.8)	16 (3.1)	4 (0.8)	0.83
problematic	SIRscan	2	25	12	0 (0)	20 (80.0)	2 (8.0)	0.01
A3								
overall	manual	64	776	95	3 (0.4)	7 (0.9)	27 (3.5)	-
overall	manual (av.)	64	97	98	1 (1.03)	0 (0.0)	1 (1.03)	0.96

Table 1: Categorization agreement between the App automatic procedure and control. The number of antibiotic reported here is the number of those for which clinical breakpoints are provided by the EUCAST[7]. The plates in sets A1 and A3 were grown on standard Mueller-Hinton (MH) growth medium, whereas we used blood enriched MH in A2.

the camera's optical distortions with a numerically generated AST image (see Supplementary S1). Moreover, while taking the picture, the application uses the device's gyroscopes and accelerometer, if available, to force the device orientation (parallel to ground, to avoid perspective distortions) and stability (to avoid motion blur). The application also displays a visual frame that helps center the Petri dish in the picture. Although Petri dishes have standard shapes (square or disk), we do not rely on this assumption for the analysis. In order to avoid numerical artifacts, perspective distortions are not corrected. Instead, we developed a simple acquisition setup which ensures parallelism between the dish and the camera's image plane (see supplementary S1).

The IP module analyzes an AST picture in three different sequential stages: plate cropping, detection of antibiotic disks and inhibition diameter measurement. The resulting diameters are used to categorize the susceptibility of the bacteria and interpret the results. These stages are described in Methods and summarized in Figure 1.

Automatic analysis performance In order to evaluate the App's performances we ran the fully automatic analysis procedure (without any manual intervention or correction) on three sets of antibiograms (A1-3), prepared in hospital routine conditions, in the university hospital of Créteil (France) and the Médecins Sans Frontières (Doctors without borders) hospital in Amman (details in Methods Table 2). The antibiogram pictures are taken with a smartphone and analyzed with the App's IP procedure. We compared results of the App on sets A1 and A2 with those of the SIRscan[15] commercial system and with manual reading (by eye with a ruler) for set A3. In the case of set A3, in order to minimize inter-operator error, we let eight technicians measure each

of the antibiograms in the set. Then, we compared the App results both with the measurements of each technician and with the average over the technicians. We define *agreement* as the exact correspondence of categorization between the App and control. Disagreement is said *very-major* if a resistance in control is reported as susceptibility by the App (R→S), *major* if a susceptibility is reported as resistance (S→R) and *minor* for any other disagreement involving an intermediate categorization (I).

The categorization results are reported in Table 1. the App shows an agreement of at least $90\,\%$ with both the SIRscan automatic system and with manual reading. Overall disagreement is low. In particular, *very major* disagreement happens with low frequency. Good agreement is also confirmed by the Cohen's kappa index calculated for each set.

the App's IP procedure proved its reliability at each step. For each data-set, the photos were automatically cropped to isolate the Petri dish. The automatic crop procedure never failed if the image respected the acquisition protocol (see Supplementary S1). The automatic pellet detection correctly found all antibiotic pellets in A1 and A2. In A3 $0.5\,\%$ of pellets where missed (false negatives). Half of the missed pellet show visible flaws (see Figure S4) and should not be considered in the analysis, according to experts advice. False positives never occurred (other objects wrongly identified as pellets). In case a pellet was missed, the users can easily add it by indicating its approximated position in the image, then a

The antibiotic labels were always correctly interpreted in A3, even if the image was not perfectly focused or the text was damaged (by bad printing or by positioning it with tweezers). In data-sets A1 and A2 the accuracy was $98\,\%$. Missclassification happened in cases of very poorly printed labels or for pellets from non-supported providers, on which the ML model was not trained. To overcome this problem, in the app we calculate a confidence value for each classification yielded by the model in order to reject missclassified labels and ask the user to identify them by eye (see Supplementary S3 for detailed results). Moreover, since the whole process is supervised by the user, missclassifications can easily be corrected.

Our new diameter measurement approach yielded good classification results over most of the available images, as showed in Table 1. Only a small fraction of the pictures, classified as *problematic* in Table 1 ($1.5\,\%$, $2.6\,\%$ and 0 in A1, A2 and A3 respectively), produced major discrepancies. Overall, diameter measurements allowed a susceptibility categorization agreement of at least $90\,\%$ for all three antibiogram sets (all images included). Details, metrics and a discussion on image contrast can be found in Methods and supplementary Table S2 and section S4.

The actual distribution of the diameter differences among manual, automatic and assisted (corrected by the user) readings of A3 are shown in Figure 2. In general, we observe that the manual measurements (done with a ruler) are on average slightly larger than the automatic and

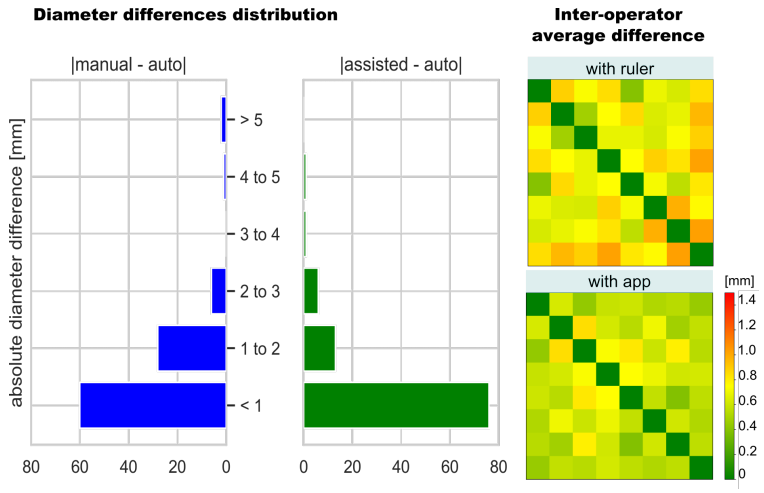

Figure 2: Visualization of the benchmark results on data-set A3. The histograms show the distribution of the absolute diameter differences between the App's automatic procedure (*auto*) and the manual measurement with ruler (blue plot) as well as with the diameter adjusted manually (*assisted*) on the smartphone by the technicians (orange plot). On the right, the heat-maps show the average absolute measurement difference among the eight technicians (given 2 readers i and j, square (i,j) represents the difference between them) measuring with the ruler (top) and in assisted mode with the App (bottom).

assisted ones. The assisted measurements are done by the technicians directly on a smartphone. The graphical interface displays a circle centered on the antibiotic disk that the user can adjust in diameter until it fits the zone of inhibition. With this kind of visualization, the measurement is easier and more accurate than the one done with a segment (which is the case of the ruler). We argue that most of the diameter difference between automatic reading and control are due to the difficulty of measuring with a ruler the diameter of inhibition zones which are not perfectly circular (see Figure S4 and Figure 4). Instead, more accurate measurements are obtained by adjusting a circular guide. The positive effect of measuring with the App is also visible in Figure 2, which displays the inter-operator average diameter difference. As expected, we observe that using the App lowered inter-operator variability.

Furthermore, we demonstrate a proof of concept for the ML classification of resistance mechanisms. We examined two clinically-relevant resistance mechanisms that are traditionally detected by the presence of non-circular inhibition zones. By training simple convolutional neural network models on the relevant portions of AST images, we obtained encouraging results: Accuracy higher than $99.7\,\%$ in detecting induction (indicative of MLSb-inducible resistant Staphylococcus aureus) and $98\,\%$ in detecting synergy indicative of ESBL production (see Supplementary S5 and S6). However, due to user experience considerations in combination with concerns about model transferability, we ultimately determined not to incorporate these resistance mechanism-detecting models into the App. We do not exclude reconsidering this approach in future version of the ap-

plication if we can generalize it to a larger and diverse data-set. Instead, every time a resistance mechanism can appear in a culture (given the bacteria species and the tested antibiotics) the application will systematically ask the user to verify the presence of the associated shape, showing illustrated examples (see Figure 3).

Finally, we compared the App to other existing systems. The categorization agreement and errors observed in this studies among the App's automatic procedure and control are similar to those of other systems (free and commercial) found in the literature (see Supplementary Table S3). The image treatment has been designed to perform on mobile devices, and does not require the user's intervention to optimize image features (e.g. contrast). We obtained consistent results even by down-scaling the antibiogram pictures up to a resolution of $1\,\mathrm{megapixel}$ (see Supplementary Table S4). The whole reading of one antibiogram ($12\,\mathrm{megapixel}$ picture, 16 antibiotic disks) takes less than $1\,\mathrm{s}$ on a PC using one $2.3\,\mathrm{GHz}$ Intel Core i5 processor, $1.5\,\mathrm{s}$ on a high-end smartphone (Pixel3 released in 2018) and $6.6\,\mathrm{s}$ on a low-end smartphone (Samsung A10), still much faster than manual reading.

Conclusions In this paper we have presented the first fully-offline smartphone application capable of analyzing disk-diffusion antibiograms. the App assists the user in taking a picture of an antibiogram, measuring and categorizing zones of inhibition, and interpreting the results. The analysis is performed entirely on the same device used to acquire the picture of the antibiogram.

Our image processing algorithm shows excellent results in all aspects of automatic AST reading (detect antibiotic pellets, read labels, measure inhibition zone diameters) and reaches an agreement of 98% with human experts. Hence, the App shows performance similar to other existing automatic reading systems. The observed accuracy is therefore considered satisfactory for usage in an AST reading system assistant. A user-friendly interface lets the user correct potential mistakes on the same device used for image acquisition.

We tested the App on antibiograms prepared with standard Mueller-Hinton (MH) growth medium as well as with MH supplemented with blood, used for fastidious organisms and obtained similar results.

We built and trained two ML-based image classification models to identify resistance mechanisms. The accuracy results are encouraging, but given the relatively small training sets, we consider the risk of over-fitting too high for the scope of such a medical device. Nevertheless, resistance mechanisms that produce visible effects on the shape of the inhibition zones are handled by the integrated expert system, which raises questions for the user to confirm/exclude the presence of such shapes.

For the evaluation of accuracy we used a $12\,\mathrm{megapixel}$ camera (Samsung A10 smartphone),

an affordable and available mobile device. Nevertheless, we observe that the resolution of the images can be degraded up to 1 megapixel and still obtain consistent results (see Table S4).

the App is now available upon request for research purposes. Further clinical investigations using the App will estimate the patient benefit enabled by AI-based antibiotic resistance testing. Pending the results, the mobile application will be released and open-sourced to the public. The fact that the App requires no hardware but a smartphone with a camera, without internet connection, makes it suitable for use in resource-limited settings, where it could also facilitate data collection. the App could also benefit many hospitals with limited resources. As a consequence, the App could help fill the digital gap and increase patients' access to AST worldwide. The use of mobile phones in infectious disease diagnostics provides a new opportunity for innovation in health systems[23]. the App is part of this mobile-Health (mHealth) revolution, which aims to increase patients' access to testing, to aid in their treatment and to decrease the digital gap among resource-rich and resource-limited settings. If integrated to a global health program, the App could facilitate the collection of epidemiological data on antimicrobial resistances, whose lack is recognized today as a major health danger[2,4]. We think that the App has the potential to help microbiology laboratories worldwide and contribute in the global fight against antibiotic resistance.

References

1. Theuretzbacher, U. *et al.* Analysis of the clinical antibacterial and antituberculosis pipeline. *The Lancet Infectious Diseases* **19**, e40 – e50 (2019).

2. de Kraker, M. E., Stewardson, A. J. & Harbarth, S. Will 10 million people die a year due to antimicrobial resistance by 2050? *PLoS medicine* **13** (2016).

3. O'Neill, J. Antimicrobial resistance: Tackling a crisis for the health and wealth of nations. *Review on Antimicrobial Resistance of the UK Department of Health* (2016).

4. De Kraker, M. E., Davey, P. G., Grundmann, H., Group, B. S. *et al.* Mortality and hospital stay associated with resistant staphylococcus aureus and escherichia coli bacteremia: estimating the burden of antibiotic resistance in europe. *PLoS medicine* **8** (2011).

5. van Belkum, A. *et al.* Innovative and rapid antimicrobial susceptibility testing systems. *Nature Reviews Microbiology* 1–13 (2020).

6. Hudzicki, J. Kirby-bauer disk diffusion susceptibility test protocol. *American Society for Microbiology* (2009). https://www.asmscience.org/content/education/protocol/protocol.3189.

7. Breakpoint tables for interpretation of mics and zone diameters. European Committee on Antimicrobial Susceptibility Testing (EUCAST) (2020). https://eucast.org/clinical_breakpoints/.

8. Performance standards for antimicrobial susceptibility testing, 30th edition. Clinical Laboratory Standards Institute (2020). https://clsi.org/.

9. Humphries, R. M. *et al.* The continued value of disk diffusion for assessing antimicrobial susceptibility in clinical laboratories: report from the clinical and laboratory standards institute methods development and standardization working group. *Journal of clinical microbiology* **56**, e00437–18 (2018).

10. Eucast disk diffusion method. European Committee on Antimicrobial Susceptibility Testing (EUCAST) (2020). https://eucast.org/ast_of_bacteria/disk_diffusion_methodology/.

11. Leclercq, R. *et al.* Eucast expert rules in antimicrobial susceptibility testing. *Clinical Microbiology and Infection* **19**, 141–160 (2013).

12. Lestari, E. S. *et al.* Comparison of the accuracy of disk diffusion zone diameters obtained by manual zone measurements to that by automated zone measurements to determine antimicrobial susceptibility. *Journal of Microbiological Methods* **75**, 177 – 181 (2008).

13. Hombach, M., Zbinden, R. & Böttger, E. C. Standardisation of disk diffusion results for antibiotic susceptibility testing using the sirscan automated zone reader. *BMC Microbiology* **13**, 225 (2013).

14. Sánchez, M., del Saz, B. S., Loza, E., Baquero, F. & Cantón, R. Evaluation of the osiris video reader system for disk diffusion susceptibility test reading. *Clinical Microbiology and Infection* **7**, 352 – 357 (2001).

15. Medeiros, A. A. & Crellin, J. Evaluation of the sirscan automated zone reader in a clinical microbiology laboratory. *Journal of Clinical Microbiology* **38**, 1688–1693 (2000). https://jcm.asm.org/content/38/4/1688.full.pdf.

16. Hejblum, G., Jarlier, V., Grosset, J. & Aurengo, A. Automated interpretation of disk diffusion antibiotic susceptibility tests with the radial profile analysis algorithm. *Journal of Clinical Microbiology* **31**, 2396–2401 (1993). https://jcm.asm.org/content/31/9/2396.full.pdf.

17. Salgado, L., Menendez, J. M., Rendon, E., Garcia, N. & Ruiz, V. Automatic antibiograms inhibition halo determination through texture and directional filtering analysis. In *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, vol. 2, 629–632 vol.2 (2001).

18. Gavoille, A., Bardy, B. & Andremont, A. Measurement of inhibition zone diameter in disk susceptibility tests by computerized image analysis. *Computers in Biology and Medicine* **24**, 179 – 188 (1994).

19. Costa, L. F. *et al.* Development of an automatic identification algorithm for antibiogram analysis. *Computers in Biology and Medicine* **67**, 104 – 115 (2015).

20. Alonso, C. *et al.* Antibiogramj: A tool for analysing images from disk diffusion tests. *Computer Methods and Programs in Biomedicine* **143**, 159 – 169 (2017).

21. Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

22. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems (2015). Software available from tensorflow.org.

23. Wood, C. S. *et al.* Taking connected mobile-health diagnostics of infectious diseases to the field. *Nature* **566**, 467–474 (2019).

24. Rother, C., Kolmogorov, V. & Blake, A. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, 309–314 (Association for Computing Machinery, New York, NY, USA, 2004).

25. Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, 2564–2571 (2011).

26. Drieux, L., Brossier, F., Sougakoff, W. & Jarlier, V. Phenotypic detection of extended-spectrum β-lactamase production in enterobacteriaceae: review and bench guide. *Clinical Microbiology and Infection* **14**, 90 – 103 (2008).

27. Kaur, J., Chopra, S., Sheevani & Mahajan, G. Modified double disc synergy test to detect esbl production in urinary isolates of escherichia coli and klebsiella pneumoniae. *Journal of clinical and diagnostic research : JCDR* **7**, 229–233 (2013).

28. Lewis, J. S. & Jorgensen, J. H. Inducible clindamycin resistance in staphylococci: should clinicians and microbiologists be concerned? *Clinical Infectious Diseases* **40**, 280–285 (2005).

29. Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37–46 (1960). https://doi.org/10.1177/001316446002000104.

30. Suzuki, S. & be, K. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing* **30**, 32 – 46 (1985).

31. Duda, R. O. & Hart, P. E. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM* **15**, 11–15 (1972).

32. Idelevich, E. A. *et al.* Evaluation of an automated system for reading and interpreting disk diffusion antimicrobial susceptibility testing of fastidious bacteria. *PLOS ONE* **11**, 1–8 (2016).

33. Google ai impact challenge competition (2019).

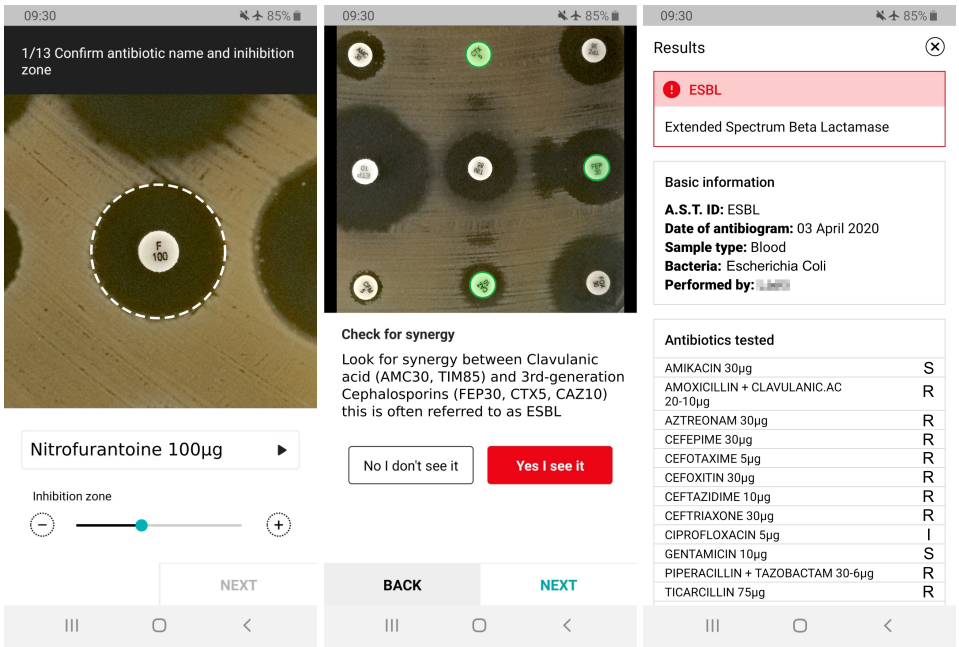

Figure 3: Screenshots of the App. From left to right: Verifying the automatic measurements on one antibiotic disk. Asking the users if they see the peculiar shape of inhibition zones that can happen in the case of a particular resistance mechanism (ESBL). Showing the final interpreted results.

Methods

The different steps of the automatic antibiogram analysis procedure implemented in the App are schematized in Figure 1. Once the antibiogram picture is taken, the Petri dish is cropped out, the antibiotic disks are found and classified according to their label, the diameter of the inhibition zones is measured and translated into susceptibility results. At this point the expert system verifies the coherence of the measurements to highlight possible errors and finally returns the interpreted results. Some screenshots of the App are shown in Figure 3 and a complete video demo can be found in the Supplementary material.

Plate cropping The first raw input to IP is an AST picture, which consists of a plate (Petri dish) to be cropped out from the remaining background (see Figure 1). Cropping is accomplished with the *GrabCut* algorithm[24], with the assumption that the plate is approximately centered in the image (i.e lies within the frame displayed on the camera screen). From the cropped plate image, we extract the dominant color, to distinguish the type of growth media (MH or blood enriched HM), and the shape of the plate (round or square). Finally, the image is converted to gray-scale for further processing.

Detection of antibiotic disks The image of a plate (Figure 1a) contains three main distinct components: the bacteria-free growth medium, the bacteria-covered growth medium, and the antibiotic disks. The latter are white round cellulose disks of known constant diameter (usually 6mm). The precision of the whole automatic analysis depends on the accuracy of measuring their position and diameter in the image. Since the disk radius is known as 6mm, the average disk radius in pixels is used to calculate the picture scale (pixel-to-mm ratio). Then, the inhibition diameters are measured from the disk center. the App features a fully-automatic method to measure antibiotic disks' positions and diameters based on intensity and shape, followed by user-assisted correction. (see Supplementary S2)

Each pellet is printed with the acronym of the antibiotic it contains. There are only a few dozen antibiotics used in AST. Nevertheless, the acronym and the print features (font, shape, size, contrast, etc.) depend on the manufacturer of the antibiotic disks. The acronym of each antibiotic disk in the analyzed antibiogram must be read in order to retrieve the corresponding breakpoint for susceptibility categorization. Previously-proposed methods for reading these acronyms compared the image moment invariants [19] or used ORB (Oriented FAST and rotated BRIEF) descriptors[20,25]. For this task, we chose ML and trained a Convolutional Neural Network (CNN) model with Tensorflow[22] (see Supplementary S3 for details). We trained the model on a total of 18,000 images of antibiotic disks from two different manufacturers (resulting in 65 unique labels) and achieved $99.97\,\%$. In order to limit the out-of-distribution error (wrong classifications of disks on which the model was not trained), we used an ensemble of 10 models and set a threshold on the output entropy. (see Supplementary S3 for details). The ML model showed to work also on poorly-printed disks and out-of-focus or low-resolution images. Interestingly we observed that even if the the printed text is damaged when the disks are placed manually (using metal tweezers), the classification is always correct.

Inhibition disk diameter measurement The inhibition zone diameter is the diameter of the largest circle centered on the antibiotic disk that does not include any bacteria; that is, the largest circle that can be drawn in the inhibition zone without touching any bacteria. In the easiest case, the inhibition forms a disk-shaped halo around the antibiotic pellet, but sometimes the disk is not well-defined, for example because of the overlap of several inhibition zones or because the antibiotic disk lies close to the plate borders (see Figure 4g). The bacteria-to-inhibition intensity contrast in the image depends on the bacteria species. Also, illumination can vary both among different images and within the same image. The observable effect is a visible difference of contrast in the AST images (see Figure 4h,i), especially when taken with a mobile phone where the illumination conditions cannot be controlled.

The algorithm for automatic diameter measurement, presented here, is referenced as SWITCH (Spatial Weighted Intensity Threshold CHangepoint). SWITCH operates a k-means clustering of

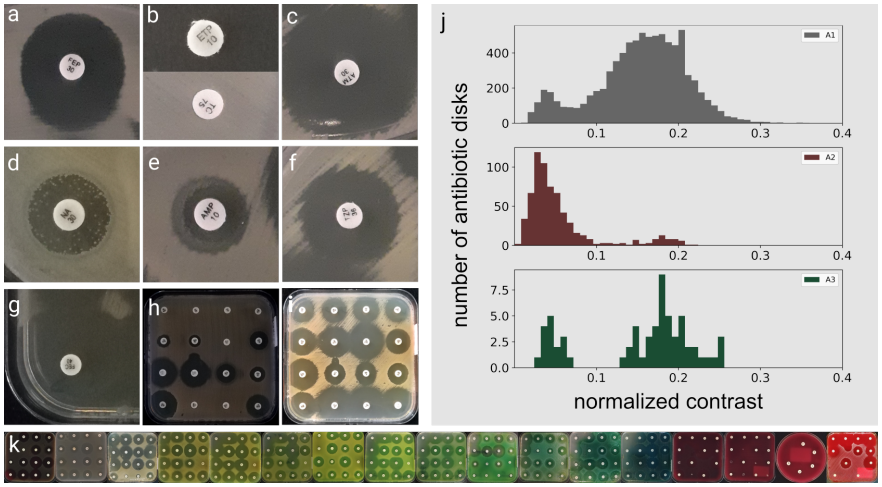

Figure 4: Variability in antibiogram pictures. Examples of difficult cases for diameter reading (a-h). Non-circular inhibition shape (a). Total or no inhibition (b). Light reflections (c). Colonies within the zone of inhibition (d). Double inhibition zones (e). Hazy borders (f). Inhibition zone overlap and plate borders (g). Low contrast (h) defined as the difference between the inhibition and bacteria intensity value, compared to a high contrast (i) image observed in A1. The histograms in (j) show the contrast variability observed in the benchmark data-sets (contrast is defined as the difference between the central gray levels of bacteria and inhibition, and normalized to the maximum available gray level). Observed variability in dominant hue (k).

the pixel intensity locally (around each antibiotic pellet) to classify inhibition and bacteria pixels (k=2). Successively, in order to find the inhibition zone boundary, it calculates and segments a radial profile $I(r)$ measured in the surroundings of the antibiotic disk (up to the closest neighboring disk). For each value of r all pixels at distance r from the pellet center are considered. The value of $I(r)$ is determined by the portion of pixels belonging to bacteria colonies (see details in Supplementary S4).

Susceptibility categorization In this study, the susceptibility categorization of the tested antibiotics (S/I/R) is done by comparing the measured inhibition zone diameters to the EUCAST clinical breakpoints[7]. The breakpoint values are stored offline in the application, within the expert system knowledge base. This base., which contains also the expert rules and other expert system resources, is maintained and updated yearly by i2a (Montpelier, France).

Auto-detection of resistance mechanisms Certain resistance mechanisms to antibiotics can be detected by disk diffusion AST because they often produce inhibition zones with characteristic shapes[26,27] (see Figure 5). These shapes appear between specific antibiotic disks. If the disks are close enough, the antibiotic molecules they diffuse can interact and produce a synergy effect against the bacteria or an induction of resistance. We used ML models to automatically recognize two particular shapes associated with two specific resistance mechanisms: synergy and induc-

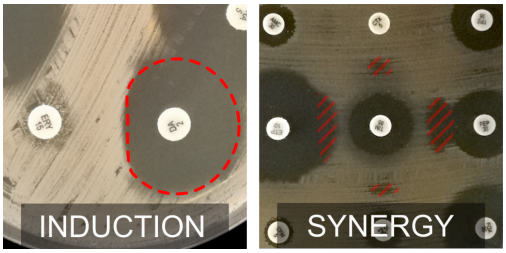

Figure 5: Characteristic shapes of inhibition zones due to resistance mechanisms.

AST set	Images	Antibiotics	Growth medium	Petri dish shape	Raters	control
A1						
overall	570	8168	MH	square	1	SIRscan
standard	561	8042	MH	square	1	SIRscan
problematic	9	126	MH	square	1	SIRscan
A2						
overall	75	649	blood	mixed	1	SIRscan
standard	73	620	blood	mixed	1	SIRscan
problematic	2	29	blood	mixed	1	SIRscan
A3	8	98	MH	circular	8	manual

Table 2: Description of the antibiogram sets used for the performance evaluation the automatic analysis procedure.

tion, which can happen with *Extended-Spectrum β-lactamase (ESBL) production*[28] and *Macrolide-inducible resistance to clindamycin* respectively.

For each of the two tests, we trained a neural network model to classify positive vs. negative images. The models to recognize Clindamycin-inducible resistance and ESBL reach an accuracy of $99.7\,\%$ and $98\,\%$, respectively (see Supplementary S5 and S6 for details). However, the models have not been proven to perform well on a wide variety of images (varying in pellet arrangement, bacteria texture, etc). Also, since classification errors can have very serious consequences in AST interpretation and patient treatment, the App would need to ask the user for confirmation when automatically detecting a resistance mechanism. So, in the best case, including the resistance mechanism models in the App would bring only modest clinical improvements. Therefore, a ML-based automatic detection procedure was not included in the current version of the mobile application. Nevertheless, the App asks users if they see such shapes wherever they are likely to appear, and shows them examples for comparison (Figure 3).

Preparation of antibiotic susceptibility tests We compared the App's automatic analysis procedure on three sets of antibiograms (A1-3) performed in hospital routine conditions, whose char-

acteristics are summarized in Table 2. The distribution of the species in these sets is reported in Supplementary Figure S7. AST groups A1 and A2 consist of 571 and 74 antibiograms prepared during working routine in the microbiology laboratory of University Hospital in Créteil, France. More than $91\,\%$ of the plates were square; the remaining $9\,\%$ were circular. The antibiotic disks were bought from i2a (Montpellier, France) and positioned onto the plate with a dispenser gun. A1 plates were prepared with standard Mueller-Hinton (MH) agar, whereas the plates in A2 contained MH supplemented with blood agar (MH-F). AST group A3 was prepared in the Hospital of Médecins Sans Frontières in Amman, Jordan. It consists of 8 plates, all circular. The antibiotic pellets were produced by Liofilchem and positioned by hand with metallic tweezers. The plates were inoculated with microorganisms purchased from the American Type Culture Collection (ATCC) and routinely used for quality control. Such strains are among the main pathogens and have known inhibition diameters to various antibiotics. A detailed description of the antibiogram sets A1-3 is reported in Supplementary S7.

Data acquisition All plates in AST groups A1 and A2 were imaged with a smartphone camera (Honor 6x with a resolution of $12\,\mathrm{megapixel}$) then the images were analyzed with the App's full automatic procedure. As control, the inhibition zone diameters of AST groups A1 and A2 were measured by a laboratory technician using a commercial automatic reading system (SIRscan[13], i2a, Montpellier, France). Since the App was still under development at the time these images were taken, we used the default Android camera application for acquisition. The diameters measured by the technicians with the SIRscan system were retrospectively extracted, as these antibiograms were performed during routine analysis in the hospital. The SIRscan system allows for correction on automatic measurements. Nevertheless, technicians did not always adjust the diameters if the adjustment did not yield a different categorization result.

Among the pictures of AST groups A1 and A2, we selected *standard* and *problematic* pictures according to the following criterion: if for more than two antibiotics in the picture we found an absolute diameter difference between the App and control values of more than $20\,\mathrm{mm}$, we considered the picture as *problematic*, otherwise it was considered *standard*. The *problematic* images are often associated with plates with defects or show very low inhibition-to-bacteria intensity contrast (due to low bacteria pigmentation and/or low illumination conditions). Nevertheless most of the low contrast images in A1 and A2 where classified as *standard* (see Figure 6). The coupled effect of bacteria pigmentation and illumination produces a considerable variability in the bacteria-to-inhibition intensity contrast. The contrast distribution in the benchmark data-sets A1-3 is represented in Figure 4.

All plates in A3 were imaged with the App on a smartphone (Samsung A10, $12\,\mathrm{megapixel}$ camera) by eight different laboratory technicians to take into account inter-operator variability (e.g. plate position, contrast and random noise, see Figure 4h,i,j). The resulting 64 images were

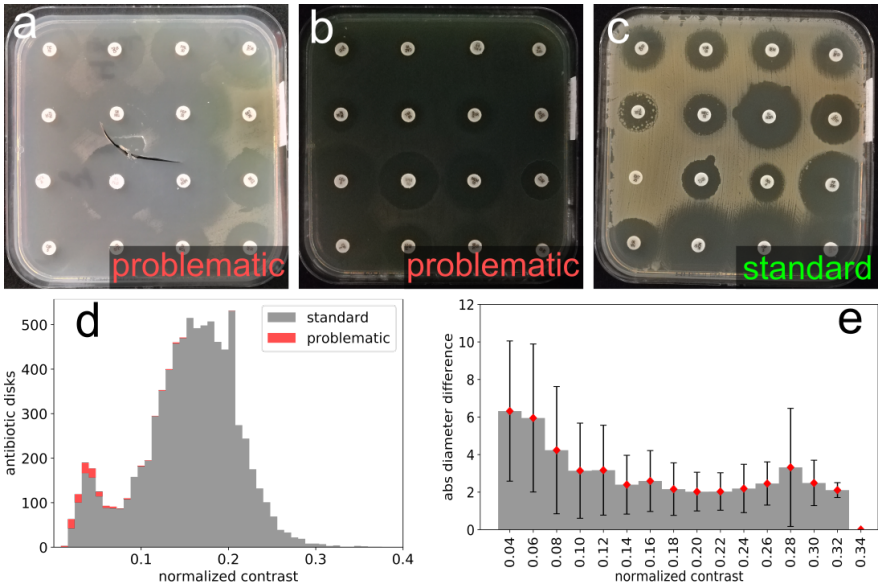

Figure 6: Images classified as *problematic*. (a) has a defect in the agar substrate and should be discarded, (b) has low contrast, (c) is a standard image for comparison. The histogram in (d) shows the distribution of image contrast for *standard* and *problematic* images in A1 (the contrast is defined here as the difference between the central intensity level of bacteria and inhibition): problematic images are mainly in the lower contrast region. Finally, (e) shows the observed mean diameter difference versus contrast: low contrast images yield worse results. Contrast is defined as the difference between the intensity value of bacteria and inhibition.

analyzed with the App 's full automatic procedure. As a control, each AST was measured manually with a ruler by the same eight lab technicians. In this way, each inhibition diameter was measured 8 times.

Benchmark The diameters of the inhibition zones read with the App's automatic procedure were compared with the control diameters. For every diameter, we calculated the absolute difference with the corresponding control value. The susceptibility categorization (SIR) of the antibiotics was made for both the App's procedure and control, by comparing the inhibition diameter of each antibiotic to the breakpoint defined in the EUCAST guidelines[7]. Antibiotics for which a breakpoint was not provided are excluded. In order to evaluate the App's performance, we compared the susceptibility categorization of the automatic procedure to the control one. Following the same terminology proposed by [16,19,20], we calculated the *agreement* as the rate of identical categorization; disagreement is classified as *very-major*, *major* and *minor*. Very-major disagreement occurs when an antibiotic is categorized S (Susceptible) while the control is R (Resistant), major error corresponds to a categorization of R with control S and minor disagreement is any other categorization error involving the Intermediate value I (Intermediate). As a measure of agreement between the App and control, we calculated the unweighted Cohen's kappa-index[29].

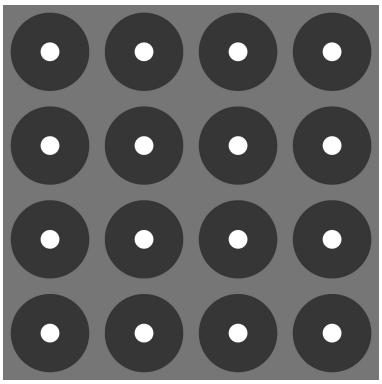

Figure S1: Numerically generated AST picture which can be printed and photographed to check if a smartphone camera is suited for analyzing antibiograms (right).

Supplementary

S1 Acquisition conditions

The precision of automatic measurements depends on image quality. In order to obtain the best possible results on images acquired with a smartphone camera, we designed a simple protocol and acquisition setup[1]. First of all, the environmental illumination should be homogeneous and strong enough to comfortably read a printed text (e.g. the pictures should not be taken in a dark room nor close to a window or table lamp). The Petri dish should be placed uncovered on a flat, black, non-reflecting surface (e.g. a piece of black felt). A dark background enhances the bacteria/inhibition contrast and is a standard recommendation in the EUCAST guidelines for antibiogram reading[10]. Place on top of the Petri dish a sheet of black cardboard held by two objects of equal height. A hole the size of the phone camera is pierced in the middle of the cardboard and the pictures are taken through the hole with the phone lying on the cardboard. The distance between the Petri dish and the smartphone (i.e. the stand height) is determined by the camera's field of view. It should be adjusted so that the Petri dish is well-fitted in the frame displayed by the camera.

The optics of smartphone cameras are not conceived for quantitative measurements, therefore small optical distortions are tolerated in production. However, initial exploration with standard calibration patters captured with several Android phones (Samsung A10, Huawei Y6, Pixel 3) suggested that optical distortion does not significantly impact measured diameters. In order for a smartphone camera to be suited for the AST measurements, the distortion it introduces should not impact the measurement more than the test sensibility (i.e. 1mm). In order to validate the smartphone camera, we generated and printed a fake antibiogram (see Figure S1). The ratio between the antibiotic disk and inhibition diameters in this picture is 25/6, so the App should measure a con-

[1] https://mpascucci.github.io/ASTapp-protocol/

stant diameter of 25mm for each pellet. This simple test allows a rapid validation of smartphones cameras before using them for AST analysis.

S2 Pellet localization and extraction.

If the antibiotic disks are placed on the antibiogram with a dispenser, their relative position is known in advance and can be exploited[19]. Unfortunately, dispensers are not always available and placing the antibiotic disks by hand remains a common practice in many laboratories, especially in resource-limited settings. Therefore, we decided not to rely on any assumptions regarding pellet position.

The detection procedure is based on the hypothesis that the pellets are white disks: their shape is approximately circular, and the intensity levels of the associated pixels in the digitized image are high. The blue channel is taken from the cropped image of the Petri dish: we observed that in pellets usually have higher contrast from the rest of the image in the blue channel. Contrast is enhanced by histogram normalization, then an intensity threshold is applied to distinguish pellets from non-pellet pixels. The threshold is 0.97 times the maximum allowed intensity level (white).

The image is de-noised with standard mathematical morphology operations. Then, the contours of the connected components are extracted [30] and filtered according to their relative size in the image and to their thinness t:

$$\mathrm{t} = \frac{l^2}{4\pi A}$$

where l is the length of the contour and A is the enclosed area. By construction, $\mathrm{t} \simeq 1$ for a contour close in shape to a circle.

This eventually allows to discard bright objects such as light speckles. The contours are considered to belong to antibiotic disks if their diameter is within 1/30 and 1/4 of the image largest size and $\mathrm{th} < 1.05$.

If an antibiotic pellet is not detected automatically, the user needs to add it manually in order to continue the analysis. The detection precision influences the further analysis results. For this task we implemented a semi-automatic procedure that takes as input the approximate coordinates of the pellet center (e.g. the user clicks on the pellet) and then accurately finds the pellet center and diameter with a Hough circle transform [31]. The previously collected information of average pellet size in the image is used to filter in the transform space.

layer type	kernel size	output size	activation
2D convolution	8	32	relu
2D convolution	4	64	relu
2D convolution	2	64	relu

Table S1: Convolution layers in the CNN model used to classify the antibiotic disk labels

S3 Antibiotic disk label reading by ML

Antibiotic name recognition is treated as a multiclass problem because there are as many codes to be recognized as there are antibiotics. This supervised learning problem is tackled via a convolutional neural network (CNN) trained using 18,000 images of antibiotic disks captured with a smartphone camera. More precisely, the images were split in a training sete of 14'000 images and a test set of 4'000 images.

The set contains 65 different labels corresponding to the pellets of two manufacturers (Liofilchem and i2a). We estimate that across the major disks manufacturers, there are over $\simeq 200$ different labels. Our model should eventually be able to identify all of those labels.

The input images are first converted to grayscale and resized to 64x64 pixels. Then, the intensity value of each pixel is normalized by subtracting the mean intensity and dividing by the intensity standard deviation. The model takes a standardized pellet image as input and returns an array of 65 values between 0 and 1 normalized to 1. These values are interpreted as the probability of belonging to a specific class (i.e. label for a given antibiotic).

The model is a CNN with three convolution layers described in Table S1. Each convolution layer is followed by max-pooling with a down-scaling factor of 2. The classification output is obtained with a random 50% dropout and a final dense layer of 65 units. We used cross-entropy as loss function.

The model is evaluated primarily on its accuracy. Since the data set doesn't cover all the antibiotic codes that could be encountered in a real use case, a second metric was introduced to evaluate the model behavior on unknown labels (hereafter "outliers"). To simulate this behavior, 5 classes (each class corresponds to a label) are held out from the dataset and the model is trained on the remaining 60. After defining a threshold on the model output, we evaluate:

- for the 60 classes from the test data
  - the accuracy on the predictions above the threshold ("accuracy on known labels")

  - the percentage of samples that fell below the threshold ("false negative rate on known labels")

- for the 5 classes held out and taken from the test data
  - the percentage of samples that fell above the threshold ("false positive rate on outliers")

A single model showed good accuracy, but a simple threshold on the output values was not good enough to minimize classification errors of outliers. Therefore, we developed an ensemble model with an entropy threshold to achieve a better tradeoff between accuracy and outlier classification error. We trained ten identical instances of this model with same hyper-parameters, same training data, but different random initialization. The models were trained using Tensorflow[22] in Python.

Classification is obtained by feeding the same pre-processed antibiotic disk image to each model of the ensemble, the ten classification outputs are averaged, and the information entropy of the average is calculated. If the entropy is larger than a fixed threshold, the argmax of the output average determines which classes the label belongs to. Otherwise, the label is considered as unknown. The threshold is chosen to optimize the trade-off between optimizing accuracy and minimizing out-of-distribution error. From the user point of view, if a disk is classified as unknown, the users must provide the label based on visual inspection. The list of possible labels are sorted according to their output prediction value, which usually brings to the top the best candidates in case of false negatives (known pellets classified as unknown).

The ensemble model achieves 99.97% accuracy on the test set. Using an ensemble of 10 models and an entropy threshold, we achieved: 100% in-distribution accuracy, 5% in-distribution false negative rate and 1% out-of-distribution false positive rate.

S4 Inhibition disk diameter measurement

Diameter measurement consists of three steps: image pre-processing, intensity radial profile extraction and segmentation.

Pre-processing The growth medium of the Petri dish is automatically determined based on the saturation, hue and intensity of the Petri dish image. First, we calculate the mean saturation, hue and intensity over the $50\,\%$ most saturated pixels in the image, then we use these values to classify the growth medium. The classifier is built with a logistic regression on more than 600 labeled images (see Figure S2) and achieves $100\,\%$ accuracy on this data set. If a red growth medium

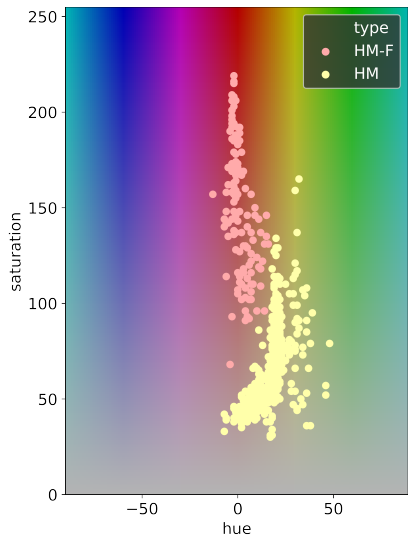

Figure S2: distribution of the mean HUE and saturation values of the AST pictures in data sets A1 and A2. Each point represents one image.

(HM-F) is detected, we use only the blue and green channels of the image when converting it to gray scale, in order to enhance the output contrast. Otherwise all three channels are used.

Successively the image is pre-processed for the diameter measurement. The Petri dish image is down-scaled to a fixed resolution of $10\,\mathrm{pixel/mm}$, which provides a reasonable trade off between precision and measurement speed. Then, all the pixels belonging to the plate borders are masked. This eliminates disturbing light reflections (due to surface tension, the agar is not flat close to the plate's borders). The plate image is segmented in Regions Of Interest (ROIs), rectangular sub-images centered on each antibiotic disk. Each ROI includes the area surrounding the disk up to the first-neighboring disks.

For the whole image first and then for each region of interest, a *k*-means classification of the pixel intensity is performed with *k*=2 (inhibition and bacteria) and the center values are stored. Since the *k*-means classification with *k*=2 is not optimal for very large inhibition zones or zones with no inhibition at all, in this case the *k*-means centers calculated on the whole image are used.

Radial profiles . The intensity of the pixels around the antibiotic disk is observed for each ROI in order to calculate a radial intensity profile $I(r)$. The profile extraction procedure takes into account both the intensity and the amount of bacteria at a given r. Without considering any privileged direction, all pixels at a given distance r from the antibiotic disk center are considered. Then $I(r)$ is given a score between 0 and 1 depending on the number n_b of "bacteria pixels". "Bacteria pixels" are defined as those pixels which have an intensity above the threshold I_{th}. The

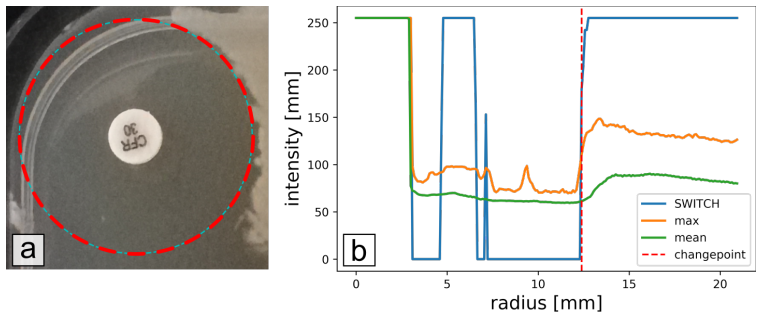

Figure S3: Radial intensity profile extraction. The region of interest of an antibiotic pellet is displayed in (a). Different radial profiles are plotted in (b) by taking the max or mean intensity value at distance r. The blue profile is calculated by the SWITCH algorithm proposed here. The peak of the signal in the inhibition zone is due to light reflection on the agar, but the segmentation correctly ignores it. In the mean profile, since the borders of the inhibition zone are not flat, the signal ramps up smoothly, introducing uncertainty in the change-point location, then the signal drops down because of inhibition zone overlap. These problems do not affect SWITCH.

bacteria intensity threshold is calculated as:

$$I_{th} = k_0 + (k_1 - k_0)(1 - s),$$

where k_0 and k_1 are the local k-means centers for inhibition and bacteria pixel intensities. s is defined as the reading sensibility. It can be fixed between 0 and 1 or automatically determined based on the image local and global contrast c_l. Contrast is defined as ($c = k_1 - k_0$), local contrast is based on local k-means (over the antibiotic disk ROI) and global contrast on global k-means (whole image).

If automatically determined, s is equal to 0.05 for very low contrast images (contrast$\leq$25) where the signal/noise ratio is low, otherwise

$$s = \frac{1}{2}\left[\left(\frac{c_g - c_l}{c_g}\right)^3 + 1\right].$$

If the local and global contrast values are similar $s \simeq 1/2$. Otherwise, it is adapted according to the ratio c_l/c_g in order to compensate noise and local intensity variability in the image.

The critical number of bacteria pixels n_b represents a spatial threshold that allows detecting the presence of bacteria at distance r even if the shape of the inhibition zone is not circular (e.g. inhibition overlap and plate borders). We choose $n_b = 1\text{mm}$ in the image scale.

Finally, $I(r)$ is segmented using a least-square fit of a step function with one degree of freedom: the change-point, which is interpreted as the inhibition zone radius in pixels. A maximum

AST set	control	images	antibiotics	a.d.d. quantiles [mm]		
				25%	50%	75%
A1						
overall	SIRscan	572	8168	0.8	1.8	3.0
standard	SIRscan	561	8042	0.8	1.8	3.0
problematic	SIRscan	9	126	2.4	6.3	20.8
A2						
overall	SIRscan	75	649	0.6	1.4	2.8
standard	SIRscan	73	620	0.6	1.4	2.6
problematic	SIRscan	2	29	11.0	17.0	22.0
A3						
overall	manual	64	784	0.0	0.8	1.4
overall	manual av.	64	98	0.1	0.6	1.3

Table S2: Absolute diameter difference (a.a.d.) between automatic measurement and control.

inhibition diameter can be specified by the user, in this case the reported diameters would be capped by that maxium (40mm by default).

The absolute diameter difference between automatic measurements and control on the test antibiotics sets are reported in Table S2. The larger absolute diameter difference in A1 and A2 can be attributed to the fact that the control diameter (measured by a commercial automatic system) was not systematically adjusted if adjustment did not have an effect on the categorization.

Published algorithms for measuring diameters Among the different algorithms for this specific task we could find in the literature[16,17] do not prove to be resilient to the common problems of non-homogeneous or non-circular inhibition zones. Same for [18], which takes into account both intensity and texture with a Student's t-test, but is sensitive to noise and assumes that the inhibition and bacteria have homogeneous textures. A recent work[19] calculates a threshold for segmentation only along 4 segments centered on the antibiotic pellet, with the risk of losing information.

S5 Resistance mechanisms: D-shape (MLSb inducible resistant Staphylococcus aureus)

The raw input to the training pipeline consists of photos of entire Petri dishes used for AST. Photos were taken using a smartphone camera at an MSF field hospital in Amman, Jordan during the course of regular AST processing. As a result, the photos represent similar conditions to those under which the App will eventually be used. Only relevant examples (those containing adjacent pairs of Clindamycin and Erythromycin pellets) are considered. The training data includes 69 positive and 143 negative examples, each image is labeled as "D-Zone or "not D-Zone.

system	licence	control	readings	agreement (%)	disagreement (%)		
					very major	major	minor
the App	free	SIRscan (A1-2)	7868	90.0	0.7	5.8	3.5
the App	free	manual (A3)	776	95.0	0.4	0.9	3.5
the App	free	manual av. (A3)	97	98.0	1.0	0	1.0
RPAA[16]	free	manual	2552	95.5	0.3	0.4	2.7
AIA[19]	free	manual	756	88.0	1.0	2.0	3.0
AntibiogramJ[20]	free	manual	720	87.0	2.0	5.0	6.0
Aura[12]	commercial	manual	4098	82.4	4.9	16.0	4.9
Osiris[14]	commercial	manual	1592	91.6	3.8	0.9	3.0
Adagio[32]	commercial	manual	1548	94.8	0.8	3.3	1.4

Table S3: Comparison of categorization agreement/disagreement of the App's performance shown in this paper versus similar studies concerning other automatic systems. This comparison should be taken with caution because the datasets used in these studies are different under many aspects (e.g. species distribution, manual measurement procedure, image acquisition and outliers dropout).

			a.d.d. quantiles [mm]			disagreement		
resolution [megapixel]	missed atb disks	unread labels	25%	50%	75%	very major	major	minor
original (12.8)	0	0	0.12	0.60	1.27	1	0	1
9.0	0	0	0.13	0.55	1.25	1	0	1
6.0	0	1	0.15	0.53	1.25	1	0	1
3.0	0	0	0.15	0.57	1.19	1	0	1
1.0	6	2	0.31	0.53	1.18	1	0	1

Table S4: Effect of resolution degradation on dataset A3 (averaged). The results are substantially the same up to resolution = 3 Megapixel. At resolution = $1\,\mathrm{megapixel}$, 6 antibiotic disks are missed out of the total 98, but the classification error remains constant.

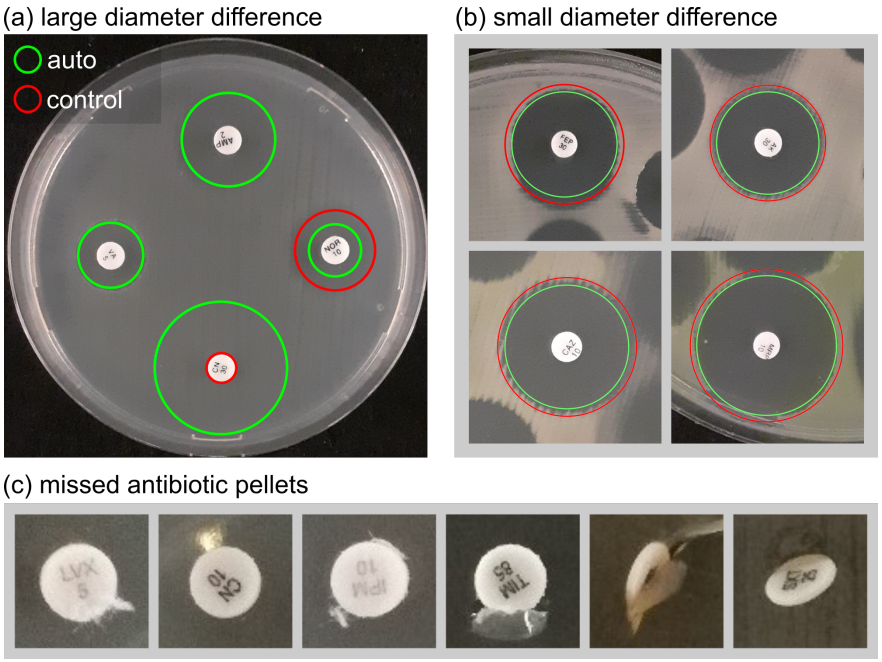

Figure S4: Automatic reading mistakes. Cases of large mistakes in inhibition diameters happen sometimes when the bacteria-inhibition contrast is low (a). Smaller diameter differences between the App and reference are probably due to the intrinsic measurement method: by visualizing a circle the App increases the measurement precision. Figure (c) shows some examples of pellets we were not able to automatically find. The three last pellets have broken the agar, hence should not be read anyways.

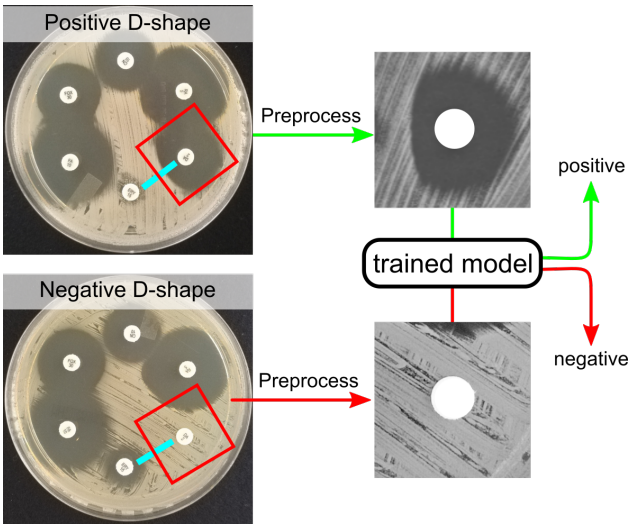

Figure S5: D-Zone prediction workflow.

First, the photos are preprocessed. The first four preprocessing steps aim to standardize input characteristics across examples (see Figure S5):

1. Identify the Clindamycin and Erythromycin pellets using the pellet label recognition strategy described above.
2. Rotate the image such that Erythromycin and Clindamycin are aligned along the x-axis, with Erythromycin on the left.
3. Crop out the biggest region surrounding Clindamycin that does not include other pellets.
4. Overlay the Clindamycin pellet with a $6\,\mathrm{mm}$ white circle in order to mask the printed text, which has no interest in the classification problem. In the interest of a small, lightweight model, the final two steps simply reduce input size without loss of model accuracy. As a result, the numerical input to the machine learning model is simply a 32x32 integer matrix, where each entry is a value between 0 (black) and 255 (white).
5. Convert the image from color to gray-scale
6. Downsize the image to $32\,\mathrm{pixel} \times 32\,\mathrm{pixel}$.

We report the model's train and test accuracy, i.e. the proportion of examples for which the model correctly predicts the known label. Since the data-set is small, a 70/30 split results in 148 training samples and 64 test samples, therefore we evaluate model structures by retraining the same model structure on multiple different train/test splits. However, the model is never tested on the same images as it was trained. This approach allows us to compare different model architectures, whereas a single train/test split would result in ties.

Models were trained using TensorFlow in Python. After experimenting with multiple neural network architectures, the architecture with the best performance has the following characteristics:

- Input size of $32\,\mathrm{pixel} \times 32\,\mathrm{pixel}$,
- 73,309 total trainable parameters,
- Binary cross-entropy loss function
- Adam optimization,
- input size of $32\,\mathrm{pixel} \times 32\,\mathrm{pixel}$,
- layers: 3x 2-D convolution layers rectified linear unit activation, with 32, 64, and 128-dimensional outputs respectively,

Data-set name	D1	D2
acquisition device	smartphone camera	SIRscan system
Number of ESBL positive examples	241	1344
Number of ESBL negative examples	181	818
Center pellet	Amoxicillin/Clavulanic acid 30 µg	Amoxicillin/Clavulanic acid 30 µg or Ticarillin/Clavulanic acid 75+10 µg

Table S5: Training data-sets for the synergy model.

- Dropout layer to avoid over-fitting,
- Densely-connected layer of size 50 with sigmoid activation (for binary classification).

Across 30 randomly-chosen train/test splits, the best model achieves on average 100% training accuracy and 99.74% test accuracy.

D-Zone detection is a relatively easy ML problem, obtaining 99.74% test accuracy with only 212 training images. The D-Zone model has not been evaluated on out-of-distribution images.

S6 Resistance mechanisms: ESBL

For the ESBL problem, we have access to two distinct data-sets (see Table S5. Each data-set consists of images of AST plates. Each image contains pellets arranged to perform a Double Disk Synergy test and each is labeled as "ESBL" or "not ESBL". The two data-sets differ highly in different aspects such as image quality and intensity contrast, bacteria culture texture, pellet arrangement, specific antibiotics used.

Before training a neural network on the images, we preprocess them to standardize across images and extract relevant regions:

1. Crop out a 35mm region surrounding amoxicillin-clavulanate. In both our data-sets, 35mm is sufficiently large to encompass amoxicillin-clavulanate and the surrounding 3rd-generation cephalosporins.
2. Convert the image to gray scale.
3. Normalize image contrast such that the intensity threshold is similar across all images.

train set (%)		test set (%)		preprocessing	accuracy (%)	
D1	D2	D1	D2		train	test
70	-	30	-	light	99.83	98.04
-	70	-	30	light	98.89	97.98
100	-	-	100	light	99.93	66.77
-	100	100	-	light	99.75	69.23
100	-	-	100	heavy	99.97	58.96
-	100	100	-	heavy	99.59	59.16

Table S6: ESBL model accuracy results.

4. Overlay the amoxicillin-clavulanate pellet with a pure black circle.

Alternative preprocessing approaches did not introduce significant improvement of the prediction accuracy. For example we tested: 1. blur the image to eliminate unnecessary details such as bacteria streaks or 2. classify each pixel as bacteria vs. inhibition using an intensity threshold computed by k-means clustering (we will refer to this strategy as "heavy preprocessing" in this section).

As in the case of D-Zone, train and test accuracy is reported. However, since in the ESBL case we have access to two distinct data-sets, we also examine model transferability. Therefore, we examine train and test accuracy within each of the following setups:

- Train and test on disjoint sets of images drawn from the same data-set
- Train and test on disjoint sets of images, where both the test and train sets are drawn from the combination of both data-sets
- Train on one data-set and test on the other

The model is over 97% accurate when its test set is drawn from the same distribution as its training set. However, when the training and test sets are of different origin, the model performs no better than random. Out-of-distribution examples are of interest because we cannot control the quality of images that the app will eventually be used to analyze. Specifically to address this problem, we tested the "heavy" preprocessing approach but did not improve cross-data-set accuracy. As more training examples are added, model accuracy increases only until 300 training examples are used. Additionally, augmenting the training data-set by transforming (e.g., rotating) images did not improve validation accuracy. Therefore, obtaining more similar data is not expected to improve results.

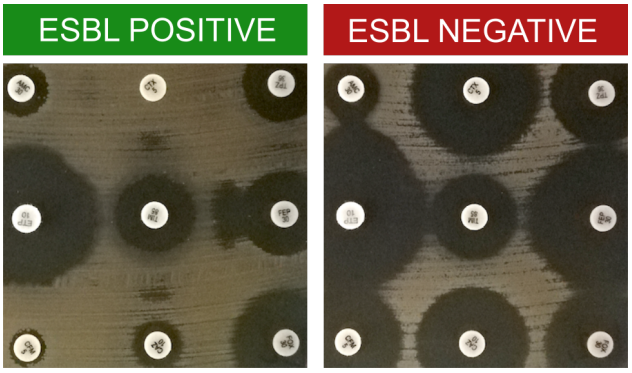

Figure S6: Examples of ESBL positive and negative images. The shapes of the inhibition zone can be very different.

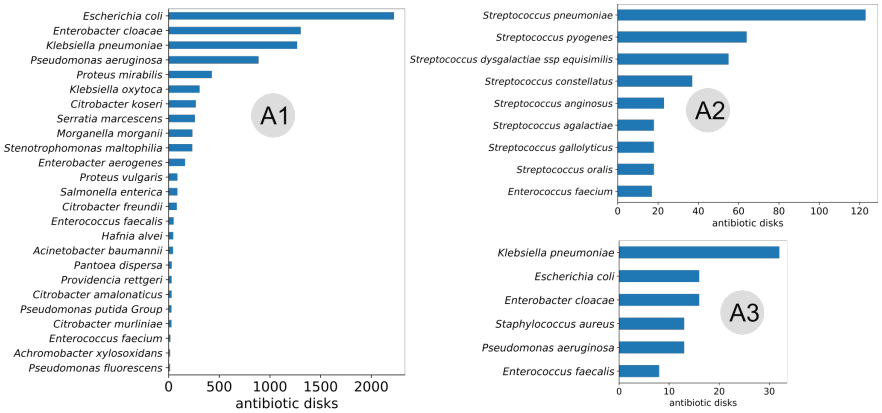

Figure S7: Distribution of species in the AST groups used for benchmarking.

ESBL classification is a much more difficult machine learning problem than D-Zone classification (see Supplementary S5). This difficulty might be due to the very large variability in the ESBL-positive examples (see Figure S6). While D-Zone examples are all very similar to each other, ESBL-positive examples may show different shapes, depending on the relative position and distance of the involved antibiotic disks, on the texture of the bacteria and on the quality of the picture.

S7 Data-sets

The bacterial species in the antibiograms sets used for benchmarking are reported in Figure S7.

A1 Antibiograms were performed according to the EUCAST[10] recommendations on Mueller-Hinton agar or MH-F agar (blood-agar) for fastidious organisms (Biorad, Marnes-la-Coquette, France).

Specifically, the eight antibiograms in A3 have been inoculated with the following ATCC dried microorganisms:

- *Pseudomonas aeruginosa* (ATCC 27583)
- *Klebsiella pneumoniae* carbapenemase producer (ATCC 700603)
- *Klebsiella pneumoniae* SHV-18-ESBL-producer (ATCC 700603)
- Fluoroquinolone susceptible *Escherichia coli* reference strains (ATCC 25922)
- Methicillin-Resistant *Staphylococcus aureus* (NCTC 12493)
- Vancomycin-sensitive *Enterococcus faecalis* (ATCC 29212)
- Gentamicin-resistant *Enterococcus faecalis* (ATCC 49532)

Acknowledgements the App was supported by: the Médecins Sans Frontières Foundation[2] , the Laboratoire de Mathématiques et Modélisation d'Évry, the Genoscope Laboratory and the bacteriology laoratory of Universitary Hospital Henri Mondor, Créteil. For its potential impact, this project was awarded the 2019 Google AI Impact Challenge, selected among over 2600 applicants worldwide[33].

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to Marco Pascucci (email: marco.pascucci@paris.msf.org).

[2]https://fondation.msf.fr/en