|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-M-008-A01 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | E-meeting, 28-30 September 2021 | |
| **DOCUMENT** | | | | |
| **Source:** | | TG-Bacteria Topic Driver | | |
| **Title:** | | Att.1 – TDD update (TG-Bacteria) [same as Meeting L] | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Nada Malou MSF France | | Email: [nada.malou@paris.msf.org](mailto:nada.malou@paris.msf.org) |

|  |  |
| --- | --- |
| **Abstract:** | This document serves as initial "skeleton" document ("DEL10.3") for the forthcoming topic description document (TDD) of the topic group Diagnosis of bacterial infection and anti-microbial resistance (AMR) (TG-Bacteria), which is concerned with the standardized benchmarking of AI for AMR-diagnosis. The outline will follow the template structure defined in FGAI4H-C-105. This TDD draft will be created in a joint effort by the topic group and continuously improved over the upcoming meetings until it is finally approved by the focus group.  This version of TDD is the same as seen in Meeting L reproduced for Meeting M for easier reference. |

**CONTENTS**

| Page |
| --- |
| [1 Introduction 4](#_Toc72149208)  [1.1 About the FG-AI4H topic group on Diagnosis of bacterial infection and anti-microbial resistance (AMR) 4](#_Toc72149209)  [1.2 Topic Description 4](#_Toc72149210)  [1.3 Relevance of the topic 5](#_Toc72149211)  [1.4 Gold standard of current health topic handling 5](#_Toc72149212)  [1.5 Possible impact of AI in this topic 5](#_Toc72149213)  [1.6 Ethical Considerations 6](#_Toc72149214)  [1.7 Existing AI Solutions 8](#_Toc72149215)  [2 Method 8](#_Toc72149216)  [2.1 Reading antibiotic wafers on a petri dish 8](#_Toc72149217)  [2.1.1 Antibiotic name recognition: AI input data structure 9](#_Toc72149218)  [2.1.2 Antibiotic name recognition: Test data labels 9](#_Toc72149219)  [2.1.3 Antibiotic name recognition: Scores & metrics 9](#_Toc72149220)  [2.1.4 Antibiotic name recognition: Benchmarking Methodology and Architecture 10](#_Toc72149221)  [2.1.5 Antibiotic name recognition: Results 10](#_Toc72149222)  [2.1.6 Antibiotic name recognition: Discussion 10](#_Toc72149223)  [2.2 Identification of resistance mechanisms: D zone (not implemented) 10](#_Toc72149224)  [2.2.1 MRSA D-test classification 10](#_Toc72149225)  [2.2.2 D-test: AI Input Data Structure 11](#_Toc72149226)  [2.2.3 D-test: AI Output Data Structure 12](#_Toc72149227)  [2.2.4 D-test: Test Data Labels 12](#_Toc72149228)  [2.2.5 D-test: Scores & Metrics 12](#_Toc72149229)  [2.2.6 D-test: Benchmarking Methodology and Architecture 13](#_Toc72149230)  [2.2.7 D-test: Results 13](#_Toc72149231)  [2.2.8 D-test: Discussion 13](#_Toc72149232)  [2.3 Identification of resistance mechanism: ESBL classification (not implemented) 13](#_Toc72149233)  [2.3.1 ESBL: AI Input Data Structure 14](#_Toc72149234)  [2.3.2 ESBL: AI Output Data Structure 16](#_Toc72149235)  [2.3.3 ESBL: Test Data Labels 16](#_Toc72149236)  [2.3.4 ESBL: Scores & Metrics 16](#_Toc72149237)  [2.3.5 ESBL: Benchmarking Methodology and Architecture 16](#_Toc72149238)  [2.3.6 ESBL: Results 16](#_Toc72149239)  [2.3.7 ESBL: Discussion 18](#_Toc72149240)  [2.4 Why complex mechanisms recognition is not in the app 19](#_Toc72149241)  [Annex A: Glossary 20](#_Toc72149242)  [Annex B: Declaration of Conflict of Interests 20](#_Toc72149243) |

**List of Tables**

| Page |
| --- |
| [Table 1: Categorization of the topic according to categorization guideline (currently C-0xx) 6](#_Toc72149244)  [Table 2: Characteristics of two distinct datasets used to train ESBL model. 14](#_Toc72149245)  [Table 3: Train & test accuracy of ESBL-detection model. To avoid idiosyncratic results of specific train/test splits, each figure is the average of 5 separate train/test splits. 16](#_Toc72149246) |

**List of Figures**

| Page |
| --- |
| [Figure 1: Example of an antibiotic wafer for Ciprofloxacin 8](#_Toc72149247)  [Figure 2: Image acquisition set up 9](#_Toc72149248)  [Figure 3: In-distribution false negative rate against out-of-distribution false positive rate for different thresholding methodologies 10](#_Toc72149249)  [Figure 4: D-shape preprocessing, training and evaluation pipeline 12](#_Toc72149250)  [Figure 5: A "difficult" D-shape example. On some train/test splits, the model incorrectly labels this image as "Not D-shape". The remaining 211 images are always correctly labeled. 13](#_Toc72149251)  [Figure 6: Light- and heavy-weight preprocessing 15](#_Toc72149252)  [Figure 6: Additional training examples over 300 do not improve validation accuracy. The same set of test images was used to evaluate models trained on a varying number of training images. 17](#_Toc72149253)  [Figure 7: In cases where the model predicts the wrong label, it generally has lower confidence than in cases where the model predicts the correct label. 17](#_Toc72149254)  [Figure 8: Higher-confidence predictions are more likely to be correct. By excluding lower-confidence predictions, the remaining predictions are up to 100% accurate.' 18](#_Toc72149255)  [Figure 9: "Difficult" ESBL examples. In the bottom row, note that ESBL-positive examples may look very different from each other. The same is true for ESBL-negative examples. 19](#_Toc72149256) |

**FG-AI4H Topic Description Document DEL10.3**

**Topic group-Bacteria**

# Introduction

This topic description document specifies the standardized benchmarking for Diagnosis of bacterial infection and anti-microbial resistance (AMR) systems. It serves as deliverable No. 10\_03 of the ITU/WHO Focus Group on AI for Health (FG-AI4H).

## About the FG-AI4H topic group on Diagnosis of bacterial infection and anti-microbial resistance (AMR)

The introduction highlights the potential of a standardized benchmarking of AI systems for Diagnosis of bacterial infection and anti-microbial resistance (AMR) to help solving important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

To develop this benchmarking framework, FG-AI4H decided to create the TG-Bacteria at the meeting "F" in Zanzibar, Tanzania, 02 – 05 September 2019.

FG-AI4H assigns a *topic driver* to each topic group (similar to a moderator) who coordinates the collaboration of all topic group members on the TDD.During FG-AI4H meeting "F" in Zanzibar, Tanzania, 02 – 05 September 2019, Nada Malou from Médecins Sans Frontiéres (MSF) was nominated as topic driver for the TG-Bacteria.

* Overview of the whole document

As part of the work of the WHO/ITU Focus Group (FG) AI for health (AI4H), this document specifies a suite of standardized benchmarking for AI-use for the interpretation of antimicrobial susceptibility testing and the identification of resistance mechanisms and patterns

The document is structured in two chapters:

* Chapter 1 introduces the topic and outlines its relevance and the potential impact that a benchmarking will have.
* Chapter 2 describes the benchmarking methodology for both reading of antibiotic discs and the identification of resistance mechanism: ESBL and D zone

The document will be updated after each discussion within ITU/WHO group

## Topic Description

Antimicrobial resistance is today recognized as one of the major public health threats. It is estimated that if no actions are taken today, 10 million deaths will be attributable to AMR by 2050[[1]](#footnote-1).

Lack of access to reliable microbiology laboratories and diagnostic tools is one of the drivers of antimicrobial resistance. This is due to several factors including: the absence of essential infrastructure, the lack of laboratory supplies and equipment's, the absence of maintenance system for equipment's and finally the lack of trained human resources in microbiology [[2]](#footnote-2). The lack of human resources is a key element and the response toward this issue includes the need for the development of simplified diagnostic tests both to perform and to interpret

Artificial intelligence can support the lack of trained human resources in the key steps of diagnostic of bacterial infection: from Gram staining reading and identification of bacteria based on their shape, to the identification of colony shapes on different culture media to finally the accurate interpretation of Antimicrobial Susceptibility Testing (Antibiogo) through the identification of resistance mechanism identified by the different shapes that can be observed on an antibiogram

This document will describe the tool that is created in order to read and interpret AST. Our project (Antibiogo) does not cover the 2 other steps of the diagnostic of bacterial infections. Antibiogo is an offline smartphone application that we aim to use for the automatic reading and interpretation of AST in LMIC where expertise in clinical microbiology is lacking

## Relevance of the topic

Increasing access to good quality microbiology tests is needed at different levels. At patient level, AMR is complicating the treatment of main infectious diseases including respiratory, urinary and sexually transmitted infection. Today, simple infections that were easily treatable becomes more and more difficult to treat and we are running out of options in term of treatments. In addition, accurate tool will ensure an accurate documentation of AMR rate especially in blind spot areas in LMIC through the implementation of national and global surveillance system (WHO, 2018). This accurate documentation will also allow the adaptation of different regulations for policy makers

## Gold standard of current health topic handling

The gold standard for correct antimicrobial susceptibility testing is the interpretation and validation performed by clinical microbiologists who is an expert in microbiology with 4 to 5 years of specialization in that field

In LMIC, such skills do not exist or are limited to national /central level laboratories. Depending of the context, AST re performed by laboratory technicians without interpretation and non-accurate results are provided to clinicians for patient treatment leading to non-rational use of antibiotic, failure of patient's treatments and bad outcome

## Possible impact of AI in this topic

In recent years, different Initiatives AI based for the improvement of the diagnostic of bacterial infections were launched [[3]](#footnote-3),[[4]](#footnote-4)

For AST, several project for the interpretation of AST were developed or are under development (references) but none of them use AI for the automatic identification of resistance mechanisms and interpretation of AST reading. The solution that we are proposing will provide for both laboratory technicians and clinicians actionable accurate interpretation that will be directly used for accurate patients' treatments allowing a better outcome but also providing reliable surveillance data for national and global system. The tool will make implementation of accurate microbiology testing easier allowing an increase in number of microbiology laboratories

Table 1: Categorization of the topic according to categorization guideline (currently C-0xx)

|  |  |  |
| --- | --- | --- |
| **Level** | | **Thematic classification** |
| Level 1 | Public Health (Level-1A) | 1.1. Health service  1.2. Health systems  1.5. Health surveillance  1.9. Communicable  sub-classes applicable:  1. Epidemiology  2. microbiology |
| Clinical Health  (Level-1B) | 1.2. Diagnosis  sub-classes applicable:  17. Infectious disease  19. Laboratory medicine |
| Level 2 (Artificial Intelligence) | | 1. Machine Learning   sub-classes applicable:  1.1. Classification |
| Level 3 (nature of data types) | | 3.2. Medical Images, photographs |
| Level 4 (origin of the data) | | 4.6 mHealth App |
| Level 5 (data collectors) | | 5.1 Service provider (technologist or doctor)  Laboratory technicians |

## Ethical Considerations

With the increasing use of AI based technology in health comes new ethical questions that are not well described yet by specific policies or recommendations.

One of the major risks identified for this project is the risk for patient privacy and confidentiality. We should differentiate the ethical considerations during the 2 different phases that our project will undergo: Study phase and routine phase

Indeed, the application will never be used in routine before an independent and accurate epidemiological study evaluating the performance of the application compared to the gold standard (the clinical microbiologist)

During the study phase, all assessment procedures will be conducted in accordance with the Declaration of Helsinki and GCP guidelines. The study and application purpose and use will be explained to each patient and Informed consent will be asked for the one enrolled in the study. Question about the possibility to take picture of their AST will be clearly asked in the consent.

We will anonymize data and we do not plan to collect any patients' personal information. Participants will be identified by a unique individual identification number that will be used to label all assessment material. Source documents and data collection forms will be maintained at the assessment site in a secure location to ensure confidentiality and by only be accessed one authorized person.

During the evaluation phase, we will establish a data safety monitoring board (DSMB). The DSMB will be established prior to any data collection as part of this study. The members of the DSMB will serve in an individual capacity and provide their expertise and recommendations.

Membership of the DSMB will consist of independent experts in antimicrobial research, infection control, diagnosis, case management and epidemiology. No independent member of the DSMB shall have any conflict of interest with the study team, the organizations funding or conducting the research, or the results of the study. The DSMB will be comprised of experts in the following areas, with an emphasis on local Jordanian expert participation:

• The study population in relevant clinical settings (Jordanian representatives)

• Clinical microbiology/infection control

• Biostatistics/data analysis

• Conduct of clinical research

The primary responsibilities of the DSMB will be to periodically review and evaluate the accumulated study data for participant safety, study conduct and progress, and 2) make recommendations to investigators concerning the continuation, modification, or termination of the study.

All discordant (major and very major) results that could lead to a different treatment for a patient will be shared and analyzed by the DSMB. If thought to be critical to optimal patient management, the discrepancy in results would be shared by the DSMB with the patient's MSF clinician.

During the study phase, we will stress all parameters identified as having a direct impact on the application performances and accuracy. At each step, in a total transparency, disclosures messages for laboratory technicians and clinicians will be available. The study was already approved by MSF Ethical Review board. The protocol is also under submission to the national Ethical board of the 3 countries where the study will be implemented

For the routine phase, we will follow the (UE) 2017/745 new regulations applicable in May 2020 that replace the COUNCIL DIRECTIVE 93/42/EEC of 14 June 1993 that defines the use of medical device

Regulation (EU) 2017/746 published on 5 May 2017 lists software in the definition of IVD-MD, among reagents and other devices. Software used as accessories (not MD but intended to be used with an MD) is also within the scope of the regulation.

In our development process, we will follow the listed requirement set by the new UE 2017/745 regulation

Other ethical considerations described in the literature are also applicable to the application that we are developing [[5]](#footnote-5). Antibiogo application will be used for the interpretation of AST but also as training tool used by the laboratory technicians to acquire step by step the interpretation expertise. Same principal will be applied to clinicians. Laboratory technicians and clinicians should be trained and guided for the use of the application and the identification of any potential ethical issue.

The development of the application takes also into consideration the variability between countries, cultures, nationalities and expertise level since the UX research methodology developed includes users from different countries, nationality and backgrounds. In addition, once the application fully developed, a qualitative assessment will be performed across 3 different countries with different user. Performances are going to be evaluated and compared across the different countries with their specificities

## Existing AI Solutions

In Antibiogo, AI is used at different stages of the AST reading and interpretation. Indeed, AI is used for the detection of antibiotic discs by reading the name of the antibiotics on the pellet in addition to the automatic recognition of specific shapes that corresponds to specific resistance mechanisms including: ESBL (Extended Spectrum Betalactamase and D zone shape for the inducible resistance to clindamycin (MLSB phenotype)

*Antibiotic name recognition*: Various applications such as AntibiogramJ[[6]](#footnote-6) and SIRScan[[7]](#footnote-7) similarly recognize antibiotic pellet names on AST images.

*Recognition of resistance mechanism: ESBL & D zone*: To our knowledge, no existing application attempts to automatically recognize ESBL or D zone for inductible clindamycine (MLSB phenotype)[[8]](#footnote-8)

# Method

## Reading antibiotic wafers on a petri dish

When performing an antibiotic susceptibility test with the Kirby-Bauer method, wafers infused with antibiotics are placed on a petri dish. Each of these wafers is marked with a code identifying the antibiotic. In order to perform an automated and/or assisted analysis of an antibiotic susceptibility test performed with the Kirby-Bauer method, the software needs to read and understand which antibiotic is present on each wafer. We trained a neural network to read pictures of wafers and achieved an accuracy of 99.97% on the test set.

A close up of a dog

Description automatically generated with low confidence

Figure 1: Example of an antibiotic wafer for Ciprofloxacin

Antibiotic codes are not standardized across the industry and are supplier dependent. There is however a fair amount of overlap between the codes used across manufacturers. We estimate the number of codes needed to be recognized for an exhaustive coverage to be about 200.

### Antibiotic name recognition: AI input data structure

We assembled a dataset of 18000 wafer pictures that were cropped out of petri dish pictures. The pictures were captured with a smartphone camera under a [controlled environment](https://mpascucci.github.io/ASTapp-overview/en/suppl/imgprotocol/) ensuring a sufficient level of standardization in terms of lighting and orientation.

A picture containing table, indoor, keyboard, desk

Description automatically generated

Figure 2: Image acquisition set up

The pictures were captured in two microbiology labs, one in Jordan (8000 images) and the other in France (10000 images) and subsequently annotated by hand. 20% of pictures were set aside to constitute a test set, the remainder was used as training set. The set contains 65 different antibiotic codes.

* + 1. **Antibiotic name recognition: AI output data structure**

The model returns an array of values between 0 and 1 and the sum of which is equal to 1, with each value representing the likelihood of classification for a specific class (i.e code for a given antibiotic).

### Antibiotic name recognition: Test data labels

Each sample in the test set is labelled with one of the 65 classes of antibiotic code. Note that to achieve the goal of exhaustive coverage, the data set will need to be extended to include the ~200 codes used by various antibiotic wafers manufacturers.

### Antibiotic name recognition: Scores & metrics

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 out-of-distribution data. To simulate this behavior, 5 classes are held out from the dataset and the model is trained on the remaining 60. After defining a threshold on the model output (such as a confidence threshold on the maximum probability in the model output), we evaluate:

* for the 60 classes from the test data
  + the accuracy on the predictions above the threshold ("in-distribution accuracy")
  + the percentage of samples that fell below the threshold ("in-distribution false negative rate")
* for the 5 classes held out and taken from the test data
  + the percentage of samples that fell above the threshold ("out-of-distribution false positive rate")

### Antibiotic name recognition: Benchmarking Methodology and Architecture

Models were trained using Tensorflow in Python via Google Cloud AI platform and evaluated using a python script.

### Antibiotic name recognition: Results

The model we trained achieves 99.97% accuracy on the test set.

Several methodologies were tested to achieve the optimal trade-off between those measures. Using an ensemble of 10 models and an entropy threshold, we achieved:

* 100% in-distribution accuracy
* 5% in-distribution false negative rate
* 1% out-of-distribution false positive rate

Chart, line chart

Description automatically generated

Figure 3: In-distribution false negative rate against out-of-distribution false positive rate for different thresholding methodologies

### Antibiotic name recognition: Discussion

The model is intended to be used as a productivity help, letting the user correct potential mistakes. The achieved accuracy is therefore considered satisfactory for this usage, although further work is required to attain an exhaustive coverage of all the wafers code used by manufacturers.

## Identification of resistance mechanisms: D zone (not implemented)

### MRSA D-test classification

Macrolide resistance in MRSA is mainly mediated by two mechanisms; MLS type B (MLSb) or efflux mechanism phenotypes. Expression of MLSb phenotype may be either constitutive or inducible in presence of low levels of inducers such as erythromycin.. Clinical failures can occur if resistance mechanisms to these drugs are inadequately tested in the laboratory. The D zone test is the major phenotypic test for the identification of such resistance mechanism. If in vitro Clindamycine seems sensitive with Erythromycin resistant, a complementary test: D zone is needed.

The presence of a D-shaped zone of inhibition between Erythromycin and Clindamycin on an AST plate indicates the presence of inducible MLSB resistance. In this case, mutants can be selected by clindamycin and clindamycin should be given back as resistant to clinicians[[9]](#footnote-9) an MRSA sample is inducibly resistant to Clindamycin. We trained a neural network model to classify positive vs. negative D-test images with 100% train accuracy and 99.6% test accuracy. [cite: [*https://www.ncbi.nlm.nih.gov/pubmed/15655748*](https://www.ncbi.nlm.nih.gov/pubmed/15655748)*]*

### D-test: AI Input Data Structure

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 during the course of regular AST processing. As a result, the photos represent similar conditions to those under which Antibiogo 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.

*Preprocessing:* Since the D-shape can only appear between a known pair of antibiotics, we can preprocess to extract the relevant region of each photo.

The first four preprocessing steps aim to standardize input characteristics across examples. Positive examples always contain a "D" shape with the flat edge on the left side of the image. Irrelevant details are omitted, such as the rotation of the pellet name.

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 a 30mm region surrounding Clindamycin. The image scale (pixel-to-mm ratio) is derived from the known physical size of the pellet.
4. Overlay the Clindamycin pellet with a 6mm white circle.

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).

1. Convert the image from color to grayscale.
2. Downsize the image to 32px x 32px.

Graphical user interface

Description automatically generated

Figure 4: D-shape preprocessing, training and evaluation pipeline

*(*[*link to drawing*](https://docs.google.com/drawings/d/1FHcFHyi9b204PwLpJCQ0bSn9JimUjSAp-jDMHTmamjk/edit)*)*

### D-test: AI Output Data Structure

Binary classification of the input image as "D-shape" or "not D-shape".

### D-test: Test Data Labels

Each image is labeled as "D-shape" or "not D-shape".

### D-test: Scores & Metrics

We report the model's train and test accuracy, i.e. the proportion of examples for which the model correctly predicts the known label.

Note:Since the dataset is small, a 70/30 split results in 148 training samples and 64 test samples. Some sets of 64 test samples result in 100% accuracy, whereas others result in 98.4% (63/64 correct), etc. Therefore, we evaluate model structures by retraining the same model structure on multiple different train/test splits. For example, if a model guessed 63/64 on one train/test split and 64/64 on another test split, its overall test accuracy would be 99.2%. This approach allows us to compare different model architectures, whereas a single train/test split would result in ties.

### D-test: Benchmarking Methodology and Architecture

Models were trained using Tensorflow in Python via Google Colab.

### D-test: Results

After experimenting with multiple neural network architectures, the architecture with the best performance has the following characteristics:

* Input size of 32px x 32px
* 73,309 total trainable parameters
* Binary crossentropy loss function
* Adam optimization
* Layers:
  + 3x 2-D convolution layers rectified linear unit activation, with 32, 64, and 128-dimensional outputs respectively
  + Dropout layer to avoid overfitting
  + 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.

A close up of a white surface

Description automatically generated with low confidence

Figure 5: A "difficult" D-shape example. On some train/test splits, the model incorrectly labels this image as "Not D-shape". The remaining 211 images are always correctly labeled.

### D-test: Discussion

D-shape detection is a relatively easy ML problem, obtaining 99.74% test accuracy (only a single sometimes-mispredicted case) with only 212 training images.

99.74% test accuracy is considered sufficient for Antibiogo's use case, especially because mobile app users will be able to override the ML model's prediction.

However, the D-shape model has not been evaluated on out-of-distribution images.

## Identification of resistance mechanism: ESBL classification (not implemented)

The Double Disk Synergy Test detects Extended-Spectrum β-lactamase (ESBL) production if there is a synergy between amoxicillin-clavalunate and 3rd-generation cephalosporins (3GC). The shape of this synergy is often described as "champagne cork" but can take other forms as well. ESBL is one of the most important resistant mechanism among enterobacteria since it inactivates third generation cephalosporin antibiotics and the only remaining choice in term of treatment is last line carbapenem antibiotics. In addition, ESBL is a threat for infection prevention control in health facilities since it is associated with hospital acquired infections. Patients infected with ESBL are usually isolated in order to avoid the spread to other non-infected patients[[10]](#footnote-10)

### ESBL: AI Input Data Structure

For the ESBL problem, we have access to two distinct datasets. Each dataset consists of images of AST plates. Each image contains pellets arranged to perform a Double Disk Synergy test.

The two datasets differ highly in terms of contrast, bacteria culture texture, pellet arrangement, specific antibiotics used, etc.

Table 2: Characteristics of two distinct datasets used to train ESBL model.

| Dataset name | Source 1 | Source 2 |
| --- | --- | --- |
| Camera | Smart phone | SIRScan machine |
| Location | MSF field hospital | French University hospital |
| Number of ESBL-Positive examples | 241 | 1344 |
| Number of ESBL-negative examples | 181 | 818 |
| Center pellet used for double-disk diffusion test | Amoxicillin/Clavulanic acid 30µg | Ticarillin/Clavulanic acid 75+10µg *or* Meropeneme 10µg |
| Example ESBL-positive image (cropped) | A picture containing black, surface  Description automatically generated |  |
| Example ESBL-negative image (cropped) | A picture containing black, tableware, accessory, enamel  Description automatically generated | A close up of a device  Description automatically generated with low confidence |

*Preprocessing:* 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-clavalunate. In both our datasets, 35mm is sufficiently large to encompass amoxicillin-clavalunate and the surrounding 3rd-generation cephalosporins.
2. Convert the image to grayscale.
3. Normalize image contrast such that the intensity threshold is similar across all images.
4. Overlay the amoxicillin-clavalunate pellet with a pure black circle.

Raw image

Lightweight preprocessing approach: crop, normalization)

Heavyweight preprocessing approach:

classify pixels as bacteria  
vs. inhibition

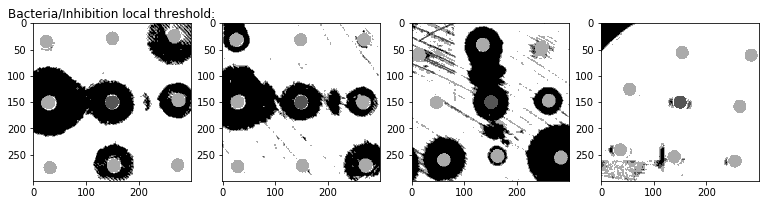
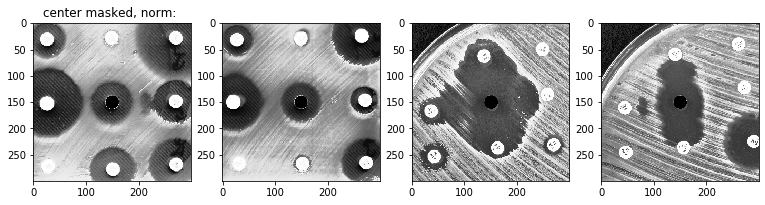


Figure 6: Light- and heavy-weight preprocessing

We also tried alternative preprocessing approaches, but these did not improve model accuracy:

* Classify each pixel as bacteria vs. inhibition using an intensity threshold computed by k-means clustering. i.e., each pixel would have input value 0=bacteria, 1=inhibition, or 2=pellet.
* Eliminate the normalization step.
* Blur the image to eliminate unnecessary details such as bacteria streaks, which could be confused for synergies.
* Overlay 3GC pellets with white circles.
* Generate additional input data by rotating and otherwise transforming the input dataset.

### ESBL: AI Output Data Structure

Binary classification: each image is labeled as "ESBL" or "not ESBL". A confidence rating can also be produced.

### ESBL: Test Data Labels

Each image is labeled as "ESBL" or "not ESBL". Recall that both positive and negative examples contain the necessary arrangement of antibiotic pellets to perform an ESBL test.

### ESBL: Scores & Metrics

As in the case of MRSA, train and test accuracy is reported. However, since in the ESBL case we have access to two distinct datasets, we also examine model transferability. Transferability between images with different characteristics is relevant to Antibiogo because images encountered in the field may differ from those used to originally train the model. 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 dataset
* Train and test on disjoint sets of images, where both the test and train sets are drawn from the combination of both datasets
* Train on one dataset and test on the other

### ESBL: Benchmarking Methodology and Architecture

Models were trained using Tensorflow in Python via Google Colab.

### ESBL: Results

Table 3: Train & test accuracy of ESBL-detection model. To avoid idiosyncratic results of specific train/test splits, each figure is the average of 5 separate train/test splits.

| Train set | Test set | Preprocessing approach | Train accuracy | Test accuracy |
| --- | --- | --- | --- | --- |
| 70% of source 1 photos | 30% of source 1 photos | lightweight: crop, mask pellet and normalize | 99.83% | 98.04% |
| 70% of source 2 photos | 30% of source 2 photos | 98.89% | 97.98% |
| 70% of source 1 + source 2 photos | 30% of source 1 + source 2 photos | 99.55%  (99.74% source 2,  99.46% source 1) | 97.44% (98.26% source 2, 96.11% source 1) |
| Source 1 photos | Source 2 photos | 99.93% | 66.77% |
| Source 1 photos | Source 2 photos | 99.75% | 69.23% |
| Source 1 photos | Source 2 photos | heavyweight: classify inhibition vs. bacteria | 99.97% | 58.96% |
| Source 2 photos | Source 1 photos | 99.59% | 59.16% |

The model is >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 to Antibiogo because we cannot control the quality of images that the app will eventually be used to analyze. To address this problem, we tried preprocessing images by classifying each pixel as bacteria vs. inhibition. However, this 'heavyweight' preprocessing approach did not improve cross-dataset accuracy.

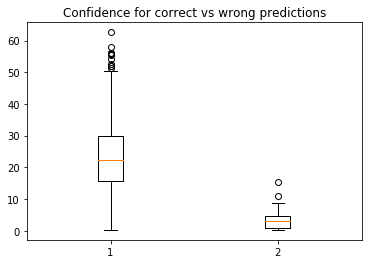
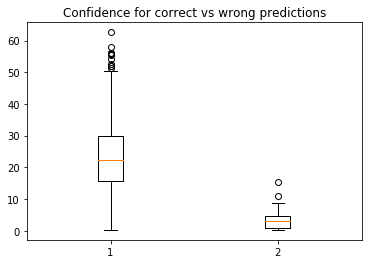
As more training examples are added, model accuracy increases only until ~300 training examples are used. Additionally, augmenting the training dataset by transforming (e.g., rotating) images did not improve validation accuracy. Therefore, obtaining more similar data is not expected to improve results.

Chart, line chart

Description automatically generated

Figure 6: Additional training examples over 300 do not improve validation accuracy. The same set of test images was used to evaluate models trained on a varying number of training images.

*Confidence thresholding:* Although the model achieved only 97% test accuracy on a mixed dataset, the model is generally less confident in its wrong guesses than its correct guesses.



Correct

predictions

Wrong

predictions

Model confidence

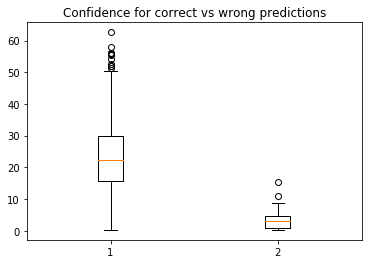


Figure 7: In cases where the model predicts the wrong label, it generally has lower confidence than in cases where the model predicts the correct label.

Therefore, we can achieve higher accuracy by excluding low-confidence cases. For example, after excluding 20% of the lowest-confidence predictions, the remaining 80% are 100% accurate. This idea could eventually be incorporated into Antibiogo e.g. by suggesting an ESBL annotation only when confidence is high.

Line chart

Description automatically generated with medium confidence

Figure 8: Higher-confidence predictions are more likely to be correct. By excluding lower-confidence predictions, the remaining predictions are up to 100% accurate.'

### ESBL: Discussion

ESBL classification is a much more difficult machine learning problem than D-shape classification. Fig. X shows that this is not due to lack of training data. Speculatively, it may be because ESBL-positive examples can take many different forms. While D-shape examples are all very similar to each other, ESBL-positive examples may show different shapes and numbers of interactions. For example, either Metropeneme or Ticarillin may be used as the center pellet for an ESBL test; but the interactions surrounding a Metropeneme look very different from those surrounding Ticaraillin. (See Fig. X)

False positives (model predicted as ESBL, but true label is not ESBL) Img911514 Img873358

False negatives (model predicted as not ESBL, but true label is ESBL)

Img909677 Img911809

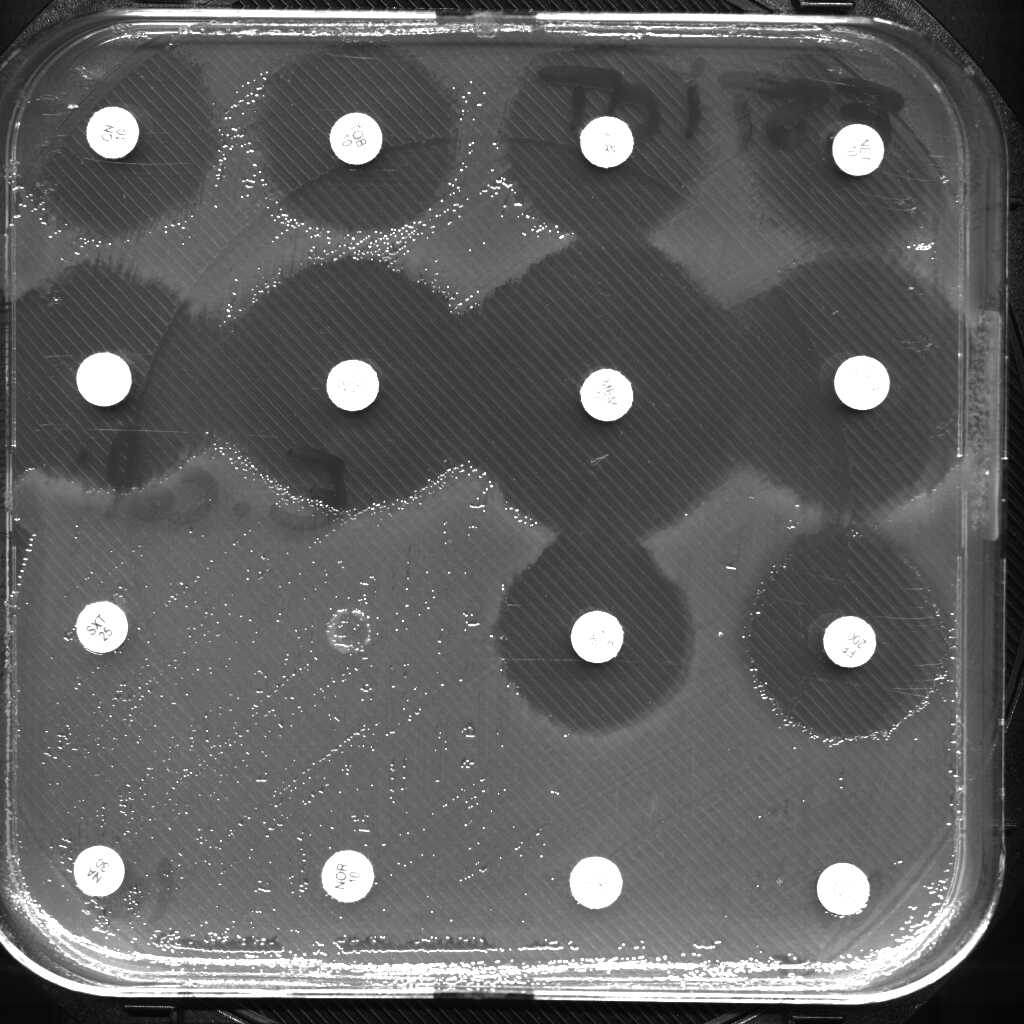
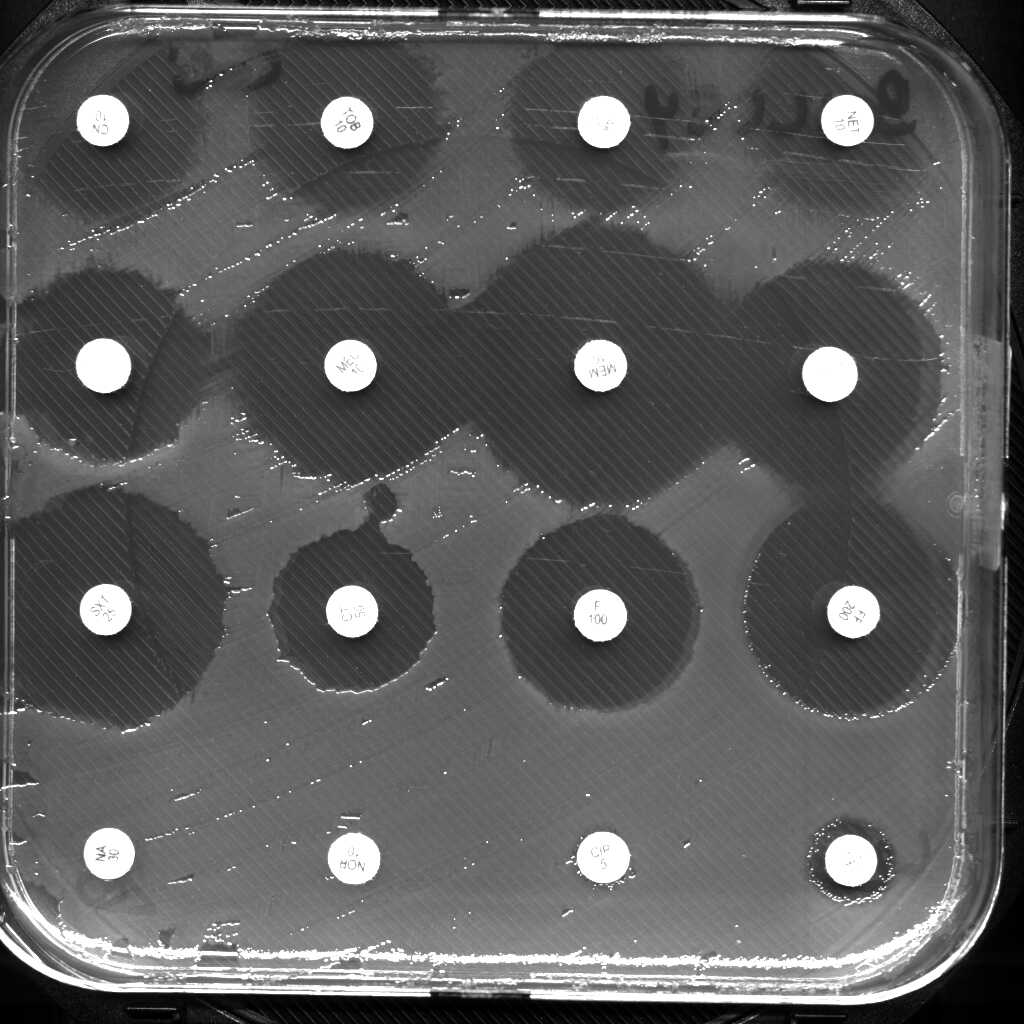
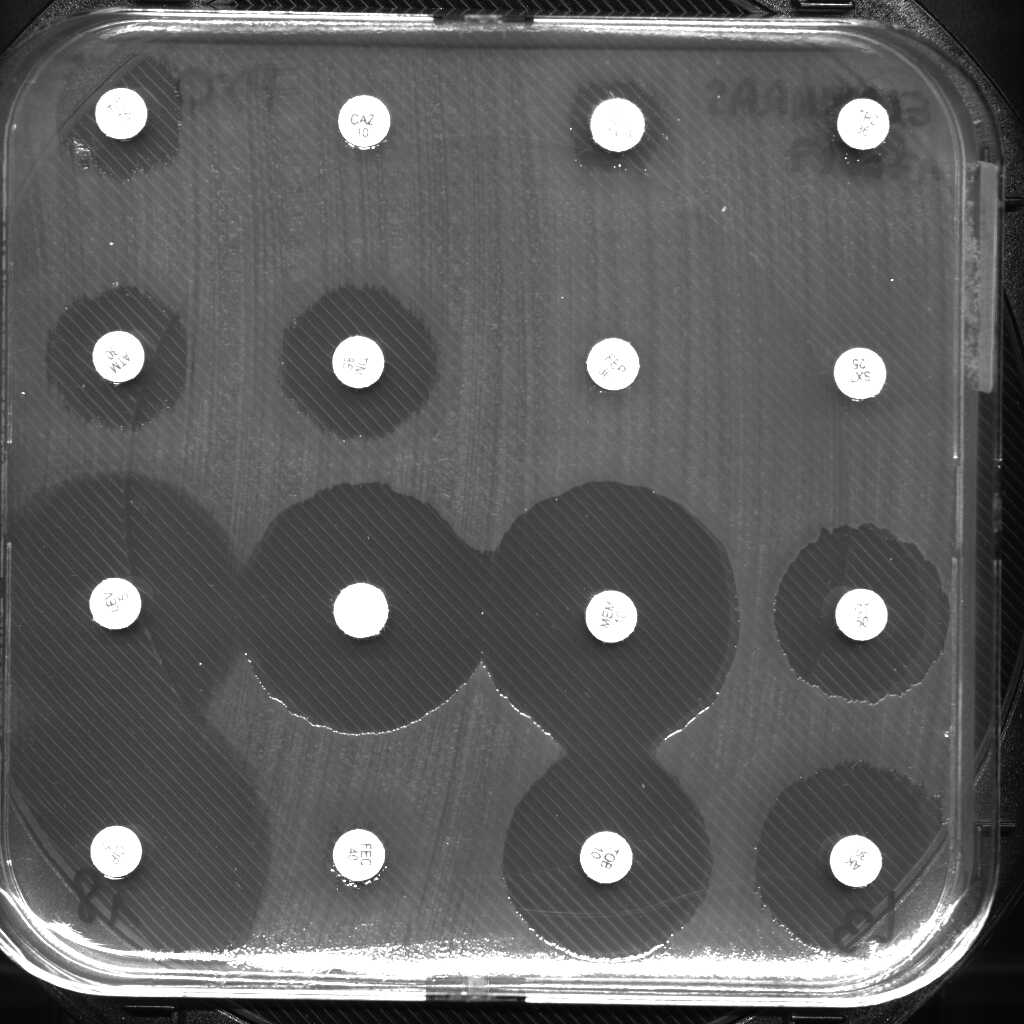
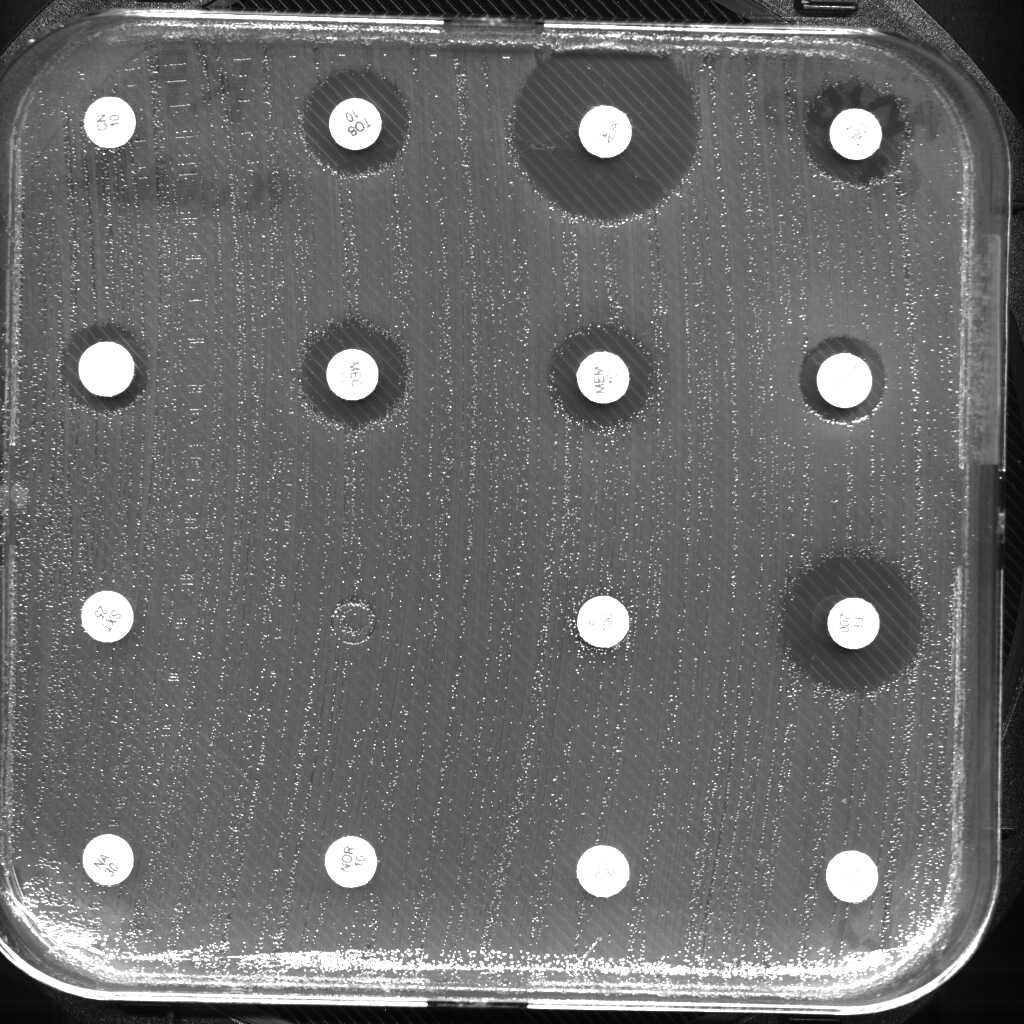
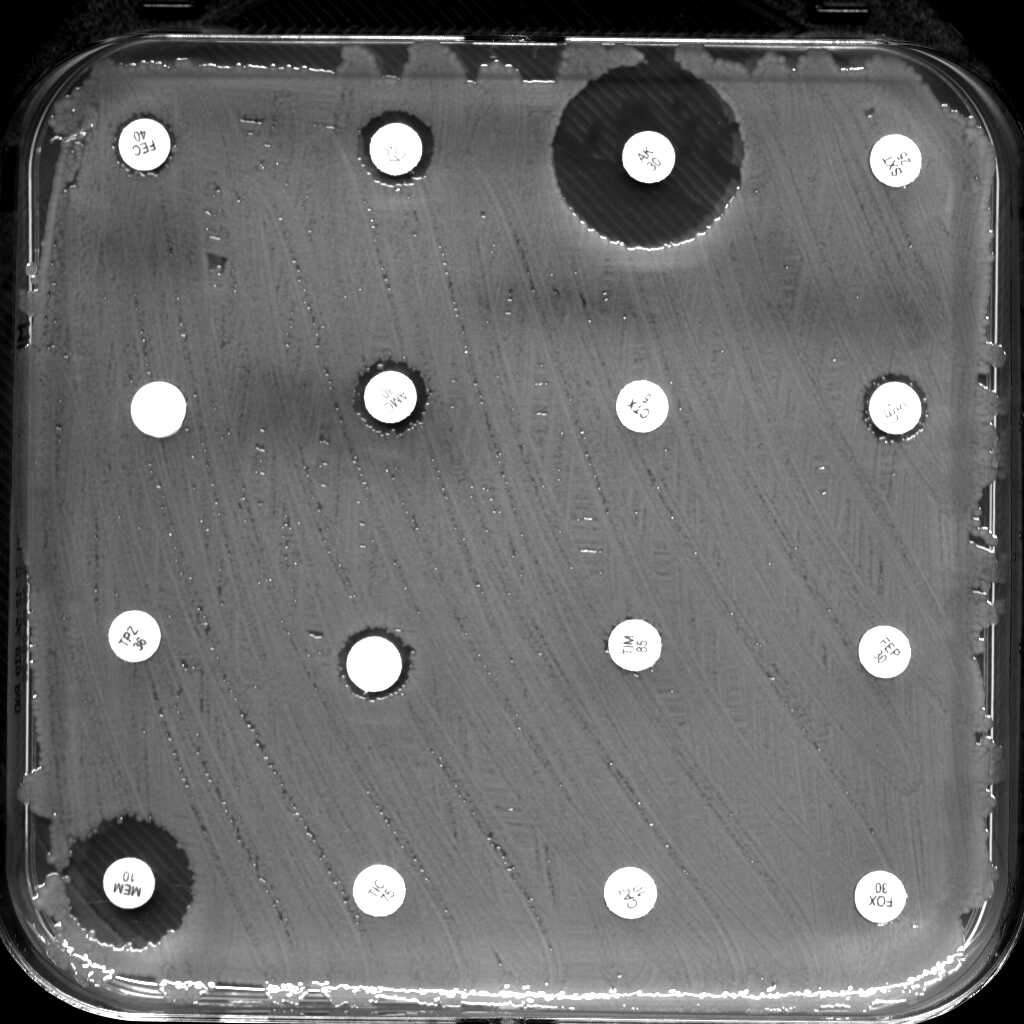
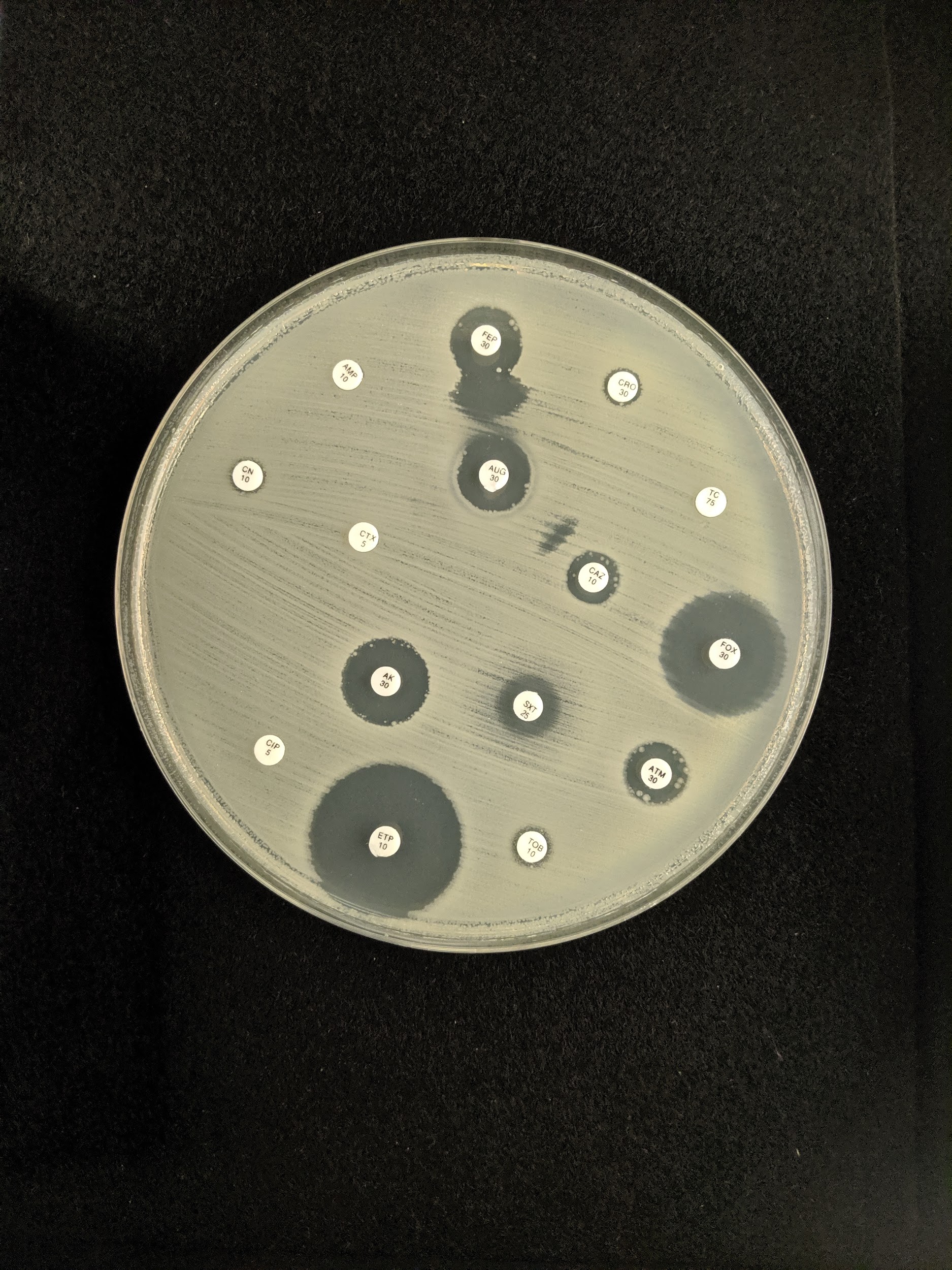
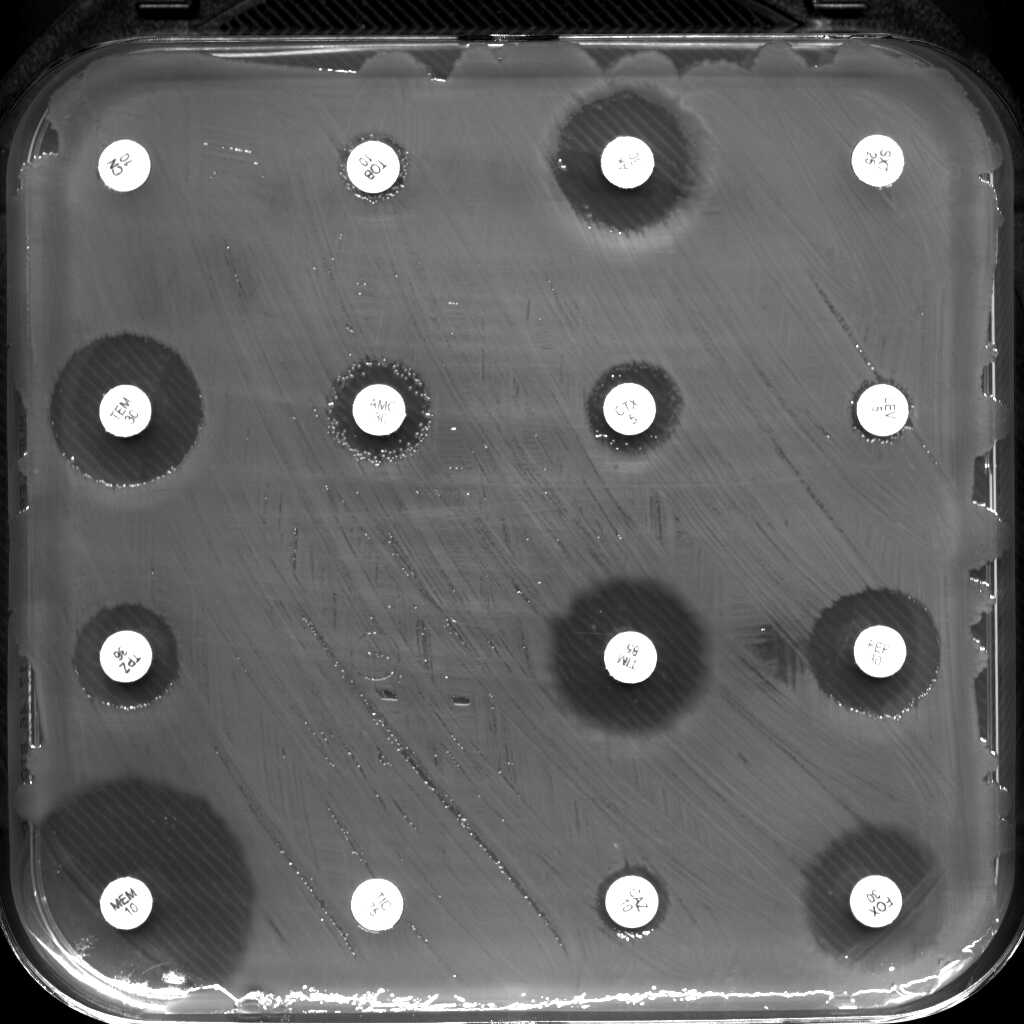
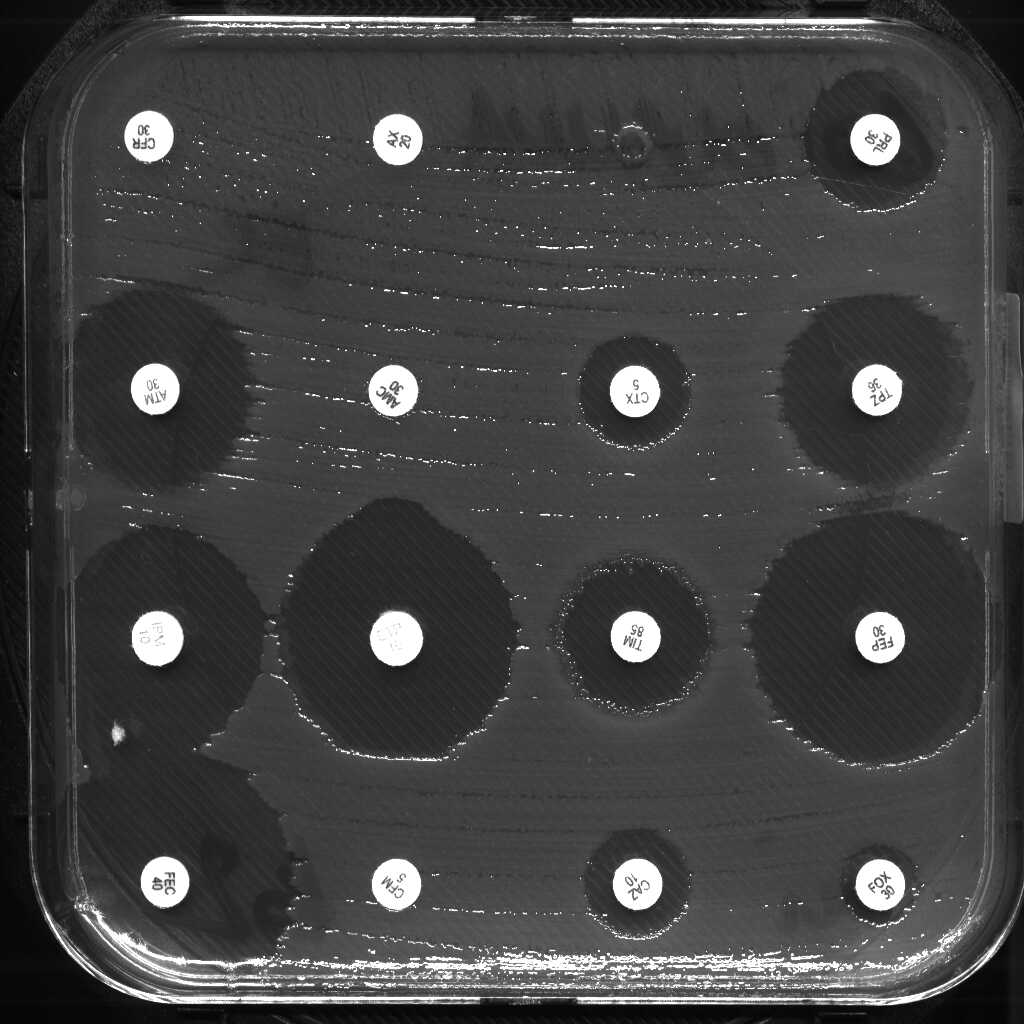


Figure 9: "Difficult" ESBL examples. In the bottom row, note that ESBL-positive examples may look very different from each other. The same is true for ESBL-negative examples.

At this stage, 97% accuracy on in-distribution images and same-as-random accuray on out-of-distribution images is not sufficient to completely rely on the ML model for final AST analysis. Further model structure optimization, preprocessing, or collection of similar training images are not expected to improve the model.

Instead, the model will provide a suggestion for the Antibiogo user to confirm. If confidence is low, no suggestion may be provided, instead relying on the user to answer the ESBL test.

Once Antibiogo is used in laboratory settings, its regular usage will enable model improvement in two ways:

1. Users will input diverse AST images, which can then be used to train the model;
2. Users will be asked either to correct the ML model's predictions, or to answer the ESBL test without ever seeing the ML prediction. Users' ESBL answers can be used as labels for further training.

To address privacy concerns, [federated machine learning](https://federated.withgoogle.com/) can retrain the model on-device, even if images and labels never leave the device.

## Why complex mechanisms recognition is not in the app

Given these models, one could imagine 3 options of how to behave when ES asks whether there is a D-zone (similar for synergy).

* **Option 1:** Run the ML model. Auto-answer "yes" or "no". Never present the question to the user.
* **Option 2**: Run the ML model. Show the user: "We think there is a D-zone. Can you confirm?"
* **Option 3**: Do not run the ML model, just ask the user "Is there a D-zone"?

Currently, the app uses Option 3. The reasons for not choosing option 1 or 2 are:

* Option 1 does not provide an opportunity for manual override. This is not acceptable for our current app requirements. End users should always have control over the app.
* The additional value of option 3 over 2 is small. The user still needs to answer a question in either case.
* According to user research, Option 2 may bias new lab techs; they would always "confirm" the suggestion, trusting the algorithm more than themselves.
* According to user research, it's easy for lab techs to detect synergy or D-zone.
* Concerns about transferability of the ML algorithm between training and real-life datasets. ML algorithm is inaccurate when trained on Dataset 1 and tested on Dataset 2, even after normalization.

# Annex A: Glossary

# Annex B: Declaration of Conflict of Interests

By each contributor to this document

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Review on Antimicrobial Resistance, O'Neill Commission [↑](#footnote-ref-1)
2. Clinical bacteriology in low-resource settings: today's solutions. Omblet et al. 2019 [↑](#footnote-ref-2)
3. Ahmed, W. M. (2013). *Classification of Bacterial Contamination Using Image Processing and Distributed Computing.*

   Bartosz Zieliński, A. P.-W. (2017). *Deep learning approach to bacterial colony classification.*

   Forero MG, C. G. (2006). *Automatic identification of Mycobacterium tuberculosis .* [↑](#footnote-ref-3)
4. Kenneth P. Smith, A. D. (n.d.). *Automated Interpretation of Blood Culture Gram Stains by Use of a Deep Convolutional Neural Network.*

   Matthew L. Faron, a. B. (n.d.). *Automatic Digital Analysis of Chromogenic Media for Vancomycin-Resistant-Enterococcus Screens Using Copan WASPLab.* [↑](#footnote-ref-4)
5. Ethical Dimensions of Using Artificial Intelligence in Health Care Michael J. Rigby. 2019. AMA Journal of Ethics® February 2019, Volume 21, Number 2: E121-124 [↑](#footnote-ref-5)
6. Antibiogramj: A tool for analysing images from disk diffusion tests. [Comput Methods Programs Biomed.](https://www.ncbi.nlm.nih.gov/pubmed/28391814) 2017 May Alonso et al. [↑](#footnote-ref-6)
7. [Standardisation of disk diffusion results for antibiotic susceptibility testing using the sirscan automated zone reader.](https://www.ncbi.nlm.nih.gov/pubmed/24099061)

   Hombach et al. BMC Microbiol. 2013 Oct 8;13:225. [↑](#footnote-ref-7)
8. Eucast expert rules and intrisec resistance: <http://www.eucast.org/expert_rules_and_intrinsic_resistance/> [↑](#footnote-ref-8)
9. Eucast expert rules: <http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Expert_Rules/2019/Staphylococcus_ExpertRules_V3.2_20190613.pdf>

   [↑](#footnote-ref-9)
10. ESBLs: A Clear and Present Danger? Rishi H.-P. Dhillon1 and John Clark. [↑](#footnote-ref-10)