Coupling machine learning and high throughput multiplex digital PCR enables accurate detection of carbapenem-resistant genes in clinical isolates

Luca Miglietta[1,2,†], Ahmad Moniri[2,†], Ivana Pennisi[1], Kenny Malpartida-Cardenas[2], Hala Abbas[3], Kerri Hill-Cawthorne[1], Frances Bolt[1], Frances Davies[1,3], Alison Holmes[1,3], Pantelis Georgiou[2] and Jesus Rodriguez-Manzano[1,*]

[1] Department of Infectious Disease, Faculty of Medicine, Imperial College London, UK.

[2] Department of Electrical and Electronic Engineering, Faculty of Engineering, Imperial College London, UK.

[3] Imperial College Healthcare NHS Trust, Hammersmith Hospital, London, UK.

[†]Luca Miglietta and Ahmad Moniri contributed equally.

**Address correspondence to this author at: Room 8N9, Commonwealth Building, Hammersmith Hospital Campus, Du Cane Road, London, W12 0NN; Email: j.rodriguez-manzano@imperial.ac.uk*

Background: The emergence and spread of carbapenemase-producing organisms (CPO) are a significant clinical and public health concern. Rapid and accurate identification of patients colonised with CPO is essential to adopt prompt prevention measures in order to reduce the risk of transmission. Recent proof-of-concept studies have demonstrated the ability to combine machine learning (ML) algorithms with real-time digital PCR (dPCR) instruments to increase classification accuracy of multiplex assays. From this, we sought to determine if this ML based methodology could accurately identify five major carbapenem-resistant genes in clinical CPO-isolates.

Methods: We collected 253 clinical isolates (including 221 CPO-positive samples) and developed a novel 5-plex assay for detection of bla_{VIM}, bla_{OXA-48}, bla_{NDM}, bla_{IMP} and bla_{KPC}. Combining the recently reported ML method 'Amplification and Melting Curve Analysis' (AMCA) with the abovementioned multiplex assay, we assessed the performance of the methodology in detecting these five carbapenem-resistant genes. The classification accuracy relies on the usage of real-time data from a single fluorescent channel and benefits from the kinetic and thermodynamic information encoded in the thousands of amplification events produced by high throughput dPCR.

Results: The 5-plex showed a lower limit of detection of 100 DNA copies per reaction for each primer set and no cross-reactivity with other carbapenemase genes. The AMCA classifier demonstrated excellent predictive performance with 99.6% (CI 97.8-99.9%) accuracy (only one misclassified sample out of the 253, with a total of 163,966 positive amplification events), which represents a 7.9% increase compared to the conventional ML-based melting curve analysis (MCA) method.

Conclusion: This work demonstrates the utility of the AMCA method to increase the throughput and performance of state-of-the-art molecular diagnostic platforms, reducing costs without any changes to instrument hardware. Our findings suggest that, pending additional validation directly from clinical samples, advanced data-driven multiplex dPCR could potentially be integrated in routine clinical diagnostic workflows.

Introduction

This paper shows that machine learning (ML) approaches coupled with high throughput real-time digital PCR (dPCR) can be used to increase detection accuracy of multiplex assays. We used a recently reported ML method called Amplification and Melting Curve Analysis (AMCA), which leverages the target-specific information encoded in each amplification event (via real-time data), to identify the nature of nucleic acid molecules (1). Validation of this methodology using clinical isolates has never been reported before. Nucleic acid amplification tests (NAATs) that incorporate the AMCA classifier for multiple target detection will greatly improve their specificity, sensitivity and turn-around times, reducing overall resource consumptions.

Antimicrobial resistance (AMR) is a serious global threat and poses a challenge for modern medicine, compromising effective infectious disease management (2,3). One of the most concerning forms of AMR is the rapid spread of carbapenemase-producing organisms (CPOs); bacteria producing enzymes that inactivate the potent antibiotics, carbapenems. Whilst overall UK incidence is low, there are centres nationally facing increasing rates and outbreaks, including Imperial College Healthcare NHS Trust (ICHNT), and it is endemic in many other regions worldwide (4,5). CPO infections are associated with higher morbidity and mortality than susceptible strains, in part because their resistance can lead to ineffective empirical therapy and suboptimal treatment (6). Therapeutic options are severely restricted, and in many cases clinical management relies on “last line” antibiotics that are less effective and have more side effects (7).

CPO infections present significant challenges for diagnostics and infection control. There is an urgent need for accurate and timely diagnosis of CPO infections to improve patient outcomes and prevent the spread of AMR. Carbapenemase resistance genes

are often co-localised on highly transmissible plasmids and are readily shared between bacterial species, providing the ideal conditions for multidrug resistant organisms. Incorrect diagnosis delays appropriate intervention, increases financial burdens for the healthcare system and complicates antimicrobial stewardship efforts. A local ICHNT economic analysis estimated the cost of a large hospital outbreak (~100 infections) of carbapenemase producing *Klebsiella pneumoniae* to be £1M. Some of the increased expenditure was associated with increased screening, bed closures, medication and patient bed-days (8).

CPO diagnostics is typically complicated and is often based upon multiple tests which utilise a wide range of instruments. Phenotypic methods typically target carbapenemase production and provide no information on the underlying resistance mechanism (9). These tests represent a low-cost (£2-15 per sample) and robust methodology; however, they rely on pure culture which increases turnaround times (12-24h) (10). A variety of molecular methods, including amplification, microarray and sequencing assays have been developed and are frequently used in microbiology laboratories (11,12). Microarray and sequencing are time consuming (>12-48h), expensive (>£50K platforms and >£80 per sample), and require bioinformatic expertise and dedicated personnel, which often limits its implementation into routine practice. Conversely, NAATs are cheaper (£18K instruments and £15-30 per sample) and faster (1-2h), however, their performance has been suboptimal for challenging diagnostic scenarios (e.g. outbreak investigation), and the application of sophisticated data processing for its optimisation (as done with the other two methods) has been largely unexplored (13).

Recently, our group has demonstrated that the large volume of data obtained from real-time digital PCR (dPCR) instruments can be exploited to perform data-driven

97 multiplexing in a single fluorescent channel, reporting a 99.28% accuracy of target
98 identification when using synthetic DNA in a 9-plex format (1). The ML method used,
99 called AMCA, leverage kinetic and thermodynamic information encoded in the
100 amplification and melting curves, to perform target identification in multiplexed
101 environments (14,15,16). Here we evaluate, for the first time, the performance of the
102 AMCA methodology for detection of the most common types of serine-beta-lactamase
103 (bla_{KPC} and bla_{OXA-48}) and metallo-beta-lactamase (bla_{IMP}, bla_{VIM} and bla_{NDM}) in clinical
104 isolates (17,18). Results were compared against another ML based classifier called
105 melting curve analysis (MCA), which only uses the thermodynamic information for
106 identification of multiple targets in a single well reaction (1,19). A 5-plex PCR assay
107 was developed in-silico and validated with synthetic DNA templates. The performance
108 of the AMCA method, using this 5-plex, was further assessed with 253 clinical isolates
109 provided by the Microbiology Department at Charing Cross Hospital, ICHNT. All
110 samples were classified in dPCR, using an intercalating dye (EvaGreen) in a single-
111 fluorescent channel. This work demonstrates that the AMCA method can be integrated
112 with conventional clinical diagnostic workflows in combination with real-time dPCR
113 platforms (Figure 1), as it does not require any hardware and only standard operation
114 modification.
115
116 **Materials and Methods**
117
118 **Synthetic DNA**
119 Double-stranded synthetic DNA (gBlock Gene fragments) containing the entire coding
120 sequences of bla_{IMP}, bla_{KPC}, bla_{NDM}, bla_{OXA-48} and bla_{VIM} was used. The gene
121 fragments (ranging from 900 to 1000 bp) were purchased from IDT and resuspended

122 in Tris−EDTA buffer to 10 ng/µL stock solutions (stored at −80 °C until further use).
123 The concentrations of all DNA stock solutions were determined using Nanodrop (Life
124 Technologies). The following NCBI accession numbers are used as reference for the
125 gBlock synthesis: NG_049172 (*bla*$_{IMP}$), NC_016846 (*bla*$_{KPC}$), NC_023908 (*bla*$_{NDM}$),
126 NG_049762 (*bla*$_{OXA-48}$) and NG_050336 (*bla*$_{VIM}$).
127

128 **Bacterial strains and culture condition**

129 A total of 253 non-duplicated *Enterobacteriaceae* isolates were collected between
130 2012-2020 from clinical or screening samples routinely processed by Microbiology
131 Department at Charing Cross Hospital, ICHNT (Ethics protocol 06/Q0406/20). Species
132 identification was performed using MALDI-TOF MS and carbapenemase mechanisms
133 were determined using the Xpert Carba-R (Cepheid) or Resist-3 O.K.N assay
134 (Corisbio). The isolates were subcultured on appropriate growth media and incubated
135 at 37 °C overnight, and the genomic DNA was extracted using GenElute Bacterial
136 Genomic DNA kit (Sigma-Aldrich) following the manufacturer's instructions.
137

138 **Primer Design**

139 The genes used in this study belong to (i) class A carbapenemase encoding for KPC
140 type, (ii) class D oxacillinases encoding OXA-48 and (iii) class B metalloenzymes
141 encoding NDM, IMP and VIM. The sequences of these genes were downloaded from
142 the GenBank web site (http://www.ncbi.nlm.nih.gov/genbank/). Based on the
143 comprehensive analyses and alignments of each carbapenemase type using the
144 MUSCLE algorithm (20), primers were specifically designed to amplify all alleles of
145 each carbapenemase gene family described above. Design and in-silico analysis were
146 conducted using GENEious Prime 2020.1.2 (https://www.geneious.com). Primer

147 characteristics were analysed through IDT OligoAnalyzer software
148 (https://eu.idtdna.com/pages/tools/oligoanalyzer) using the J. SantaLucia
149 thermodynamic table for melting temperature (T_m) evaluation, hairpin, self-dimer, and
150 cross-primer formation (21). The T_m of the amplification product of each gene was
151 determined by Melting Curve Predictions Software (uMELT) package (https://dna-
152 utah.org/umelt/umelt.html). To confirm the specificity of the real-time digital PCR
153 assays, the primers were evaluated in a singleplex PCR to ensure that they correctly
154 amplified their respective loci and that the amplicons showed the expected T_m. All
155 primers were synthesized by IDT (Coralville, IA, USA). Primer sequences and
156 amplicon information are listed in Table 1.
157
158 **Multiplex real-time digital PCR**
159 Each amplification reaction was performed in 4 µL of final volume with 2 µL of SsoFast
160 EvaGreen Supermix with Low ROX (BioRad, UK), 0.4 µL of 20X GE Sample Loading
161 Reagent (Fluidigm PN 85000746), 0.2 µL of PCR grade water, 0.2 µL of 20X multiplex
162 PCR primer mixture containing the three primer sets (10 µM of each primer), and 1.2
163 µL of different concentrations of synthetic DNA (or controls). PCR amplifications
164 consisted of a hot start step for 10 min at 95°C, followed by 45 cycles at 95°C for 20
165 s, 67°C for 45 s, and 72°C for 30 s. Melting curve analysis was performed with one
166 cycle at 65°C for 3 s and reading from 65 to 97°C with an increment of 0.5°C. We used
167 the integrated fluidic circuit controller to prime and load qdPCR 37K digital chips and
168 Fluidigm's Biomark HD system to perform the dPCR experiments, following
169 manufacturer's instructions. More specifically, each digital chip contains 48 inlets,
170 where each inlet is connected a panel consisting of 770 wells (0.85 nL well volume)
171 (22). In this study, we used 7 digital chips, totalling 336 panels.

172

173 **Analytical sensitivity for the 5-plex assay**

174 Limit-of-detection (LOD) for each primer set in the multiplex format was evaluated with
175 10-fold dilutions of synthetic DNA containing the 5 carbapenemase genes, ranging
176 from 10^1 to 10^7 copies per reaction. Each experimental condition was run in triplicate
177 using a qPCR platform (LightCycler 96, LC96). Data was analysed using LC96 System
178 software version SW1.1 (Supplemental Figure 1).

179

180 **ML-based MCA and AMCA methods**

181 The proposed method, referred to as AMCA, trains a supervised machine learning
182 model to combine the predictions of ACA and MCA as reported in previous studies
183 from Moniri et al. (1,14). In this study, the MCA method consists of applying a logistic
184 regression model to melting T_m values extracted from each melting curve. The ACA
185 consists of applying a k-nearest neighbors (KNN) model (with parameter k=10) to the
186 entire real-time curve from each amplification event. Both MCA and ACA output 5
187 probabilities associated with each target in the 5-plex. Therefore, these probabilities
188 are concatenated into 10 values which are the input to the AMCA method (which is a
189 linear regression model). It is important to note that this classifier is tuned with its own
190 cross-validation step in order to avoid overfitting.

191

192 **Statistical Analysis**

193 Prior to conducting this study, a sufficient number of samples was determined to
194 provide statistically significant results via the binomial proportion confidence interval
195 method (23). Under the assumption that the test has a sensitivity and specificity of

196 95% with a 5% margin of error, the number of samples were determined as 72 (which
197 is significantly smaller than 221 used in this study).
198
199 Performance of the models was determined by training on synthetic DNA and testing
200 on the aforementioned 221 isolates. Alongside evaluating the performance of the
201 models on the isolates, performance of the models was also estimated based on out-
202 of-sample classification accuracy, as determined by 10-fold cross-validation on the
203 training data (using stratified splits). The two-sided t-test with unknown variances was
204 used to determine statistical significance for comparing the classification accuracy of
205 AMCA against MCA. Prior to this test, a Lilliefors test was used to determine normality
206 of the distributions and the Bartlett test for equal/unequal variances. A *p*-value of 0.05
207 was used as a threshold for statistical significance for all tests.
208
209 **Results**
210
211 **Primer characterisation for optimal multiplex assay performance**
212 (i) *in-silico analysis*. To test the inclusivity and exclusivity of the 5-plex assay, primers
213 were subjected to a general NCBI BlastN search against more than 500 sequences
214 per target. Inclusivity results showed over 99% identity coverage for each target
215 (inclusivity alignments are provided in Supplemental Figure 2-6). For exclusivity
216 analysis, BlastN hits with an identity score lower than 80% were regarded as negative.
217 No cross-reactivity was observed with other sequences deposited in the database.
218 (ii) *experimental results in qPCR*. The 5-plex has been validated using a conventional
219 qPCR platform with synthetic DNA templates at concentrations ranging from 10^2 to
220 10^7 DNA copies/reaction. Supplemental Figure 1 shows the real-time amplification (a),

221 melting (b) and standard curves (c). The amplification and melting curves have distinct
222 shape for each target, which is beneficial for AMCA classification, and the observed
223 T_m values for *bla*$_{IMP}$, *bla*$_{KPC}$, *bla*$_{NDM}$, *bla*$_{OXA-48}$ and *bla*$_{VIM}$ are 81.4, 89.5, 90.2, 83.8 and
224 87.9°C, respectively. Moreover, each primer set (in a multiplex environment) shows
225 an excellent LOD of 100 DNA copies/reaction. Corresponding standard curves,
226 illustrating the C_q value as a function of the target concentration, yield an assay
227 efficiency of 87.3, 103.5, 105.7, 98.7, 88.1%, respectively.
228 (iii) *experimental results in dPCR*. The 5-plex was further validated in dPCR platform
229 with synthetic DNA templates at concentrations ranging from 10^2 to 10^7 DNA
230 copies/reaction, as depicted in Figure 2. A total of 29,165 amplification reactions were
231 performed. In Figure 2, a and b show respectively the amplification and melting curves
232 for the five carbapenem-resistant genes and the average characteristic shape for each
233 target (black solid line). Figure 2c represents the distribution of melting temperature in
234 real-time dPCR, where the T_m range for each target is computed as: *bla*$_{IMP}$ (81.3,
235 83.2˚C), *bla*$_{KPC}$ (89.0, 91.5˚C), *bla*$_{NDM}$ (90.0, 92.7˚C), *bla*$_{OXA-48}$ (83.7, 86.6˚C) and
236 *bla*$_{VIM}$ (87.7, 90.8˚C).
237
238 **Samples**
239 As depicted in Table 2, the 253 pure bacterial strains were identified from MALDI-TOF
240 MS as *Acinetobacter* spp. (n = 2), *Citrobacter* spp. (n = 16), *Enterobacter* spp. (n =
241 37), *Escherichia* spp. (n = 57), *Klebsiell*a spp. (n = 133), *Proteu*s sp. (n = 1),
242 *Pseudomonas* sp. (n = 5) and *Serratia* sp. (n = 2). Carbapenemase genes were
243 determined as a single enzyme in 220 strains (*bla*$_{IMP}$ = 45; *bla*$_{KPC}$ = 9; *bla*$_{NDM}$ = 74;
244 *bla*$_{OXA-48}$ = 84; *bla*$_{VIM}$ = 8) and as a combination in one isolate (*bla*$_{NDM}$ & *bla*$_{OXA-48}$).
245 Thirty-two isolates were confirmed as negative for the five carbapenemase genes. A

246 more detailed description of each isolate, including bacterial species, date of sampling,
247 specimen type and antibiotic resistance mechanisms can be found in Supplemental
248 Table 1.
249
250 **The AMCA model: training and cross-validation**
251 Our study aims to validate the performance of the AMCA method for detection of
252 carbapenem-resistant genes in clinical isolates. Results obtained from AMCA were
253 compared with the MCA approach. To train both models, a total of 99,860 amplification
254 events were generated using synthetic DNA templates, of which 29,165 were positive:
255 bla_{IMP} (N=4,227), bla_{KPC} (N=3,993) bla_{NDM} (N=3,500), bla_{OXA-48} (N=3,771) and bla_{VIM}
256 (N=4,596). Training dataset performance is shown in Supplemental Figure 7 with
257 relative accuracies and sample sizes.
258
259 **The AMCA model: clinical validation**
260 A total of 253 clinical isolates, including 221 positives, and 227,150 amplification
261 events (of which 163,966 positives) were used for the clinical validation. The overall
262 observed accuracy for MCA was 91.7% (CI 87.59 to 94.79%) and 99.6% (CI 97.82%
263 to 99.99%) for AMCA, which represent a 7.9% increase. A total of 21 samples were
264 misclassified for the MCA method and considered false positives (FP) as shown in
265 Table 3, whereas the AMCA reduced the number of misclassified samples to 1 (Table
266 4). All the false positive samples were identified as double infection because of the
267 overlapping distribution in the melting T_m, as shown in Figure 2. Performance
268 improvement in the AMCA method is due to the addition of real-time amplification data,
269 contrary to the MCA approach that only takes into account the melting curve
270 distribution (Supplemental Figure 8). Moreover, 32 bacterial isolates not carrying the

271 five carbapenemase genes were used to evaluate the assay specificity. The 5-plex
272 assay showed negative results in the absence of the specific target.
273
274 **Discussion**
275
276 In the last decade, novel pandemic outbreaks and the continued threats of emerging
277 multi-drug resistant microorganisms have increased the requirement of PCR tests for
278 diagnosis. Therefore, alternative methods are necessary to accelerate infection
279 source identification, especially in countries where the number of laboratory
280 capabilities is limited. The AMCA technology is designed to provide a high-throughput
281 and cost-effective testing for a large range of urgent antibiotic resistance threats in a
282 single well reaction. Seamlessly integrated with conventional diagnostic workflows,
283 this machine learning based approach can enhance multiplexing capabilities of
284 traditional qPCR and dPCR instruments, which already existing in laboratories around
285 the world. In this paper, we performed a clinical validation of the AMCA methodology
286 by targeting the "big five" carbapenem-resistant gene in multiplex PCR: bla_{VIM}, bla_{OXA-}
287 $_{48}$, bla_{NDM}, bla_{IMP} and bla_{KPC}. The performance was evaluated in both real-time PCR
288 and digital PCR instruments, and further investigated through the identification of 253
289 clinical isolates form patients affected by different pathogenic species. We found 221
290 CPO-positive samples for the mentioned targets across multiple bacterial species.
291 As previously reported (1), the AMCA method is capable of enhancing high-level
292 multiplexing capabilities in real-time dPCR platforms, increasing the classification
293 accuracy by utilising kinetic and thermodynamic information of the amplification event.
294 The methodology is extremely advantageous to identify multiple targets in a single
295 reaction in a time and cost-effective manner. In this study, we compared the

296 classification performance of two machine learning based methods, MCA and AMCA,
297 using five carbapenem-resistant genes. Our objective was to evaluate the benefit of
298 advance data analytics applied to conventional diagnostic workflows for AMR. In the
299 first instance, we addressed the MCA method performance; it can be observed that
300 this approach is limited by the thermodynamic nature of the reaction, which negatively
301 influences both sensitivity and accuracy of the multiplex assay. The visualisation of
302 the melting peak distribution showed the presence of overlapping regions (e.g. bla_{NDM}
303 and bla_{KPC}). This phenomenon can compromise the classification based purely on
304 melting curves, with further complication in double-infections events. However, this
305 effect is reduced with AMCA method as it benefits from the kinetic information of the
306 amplification event, which makes the method more reliable and reduces the need for
307 lengthy assay optimisation.
308 Generally, digital PCR offers a number of clear advantages compared to qPCR,
309 representing an attractive alternative tool for molecular diagnostics (24). Due to its
310 high throughput and the single molecule environment, dPCR is extremely suitable for
311 machine learning based algorithm and high-level multiplex assays, which further
312 increases the diagnostic capabilities and enables multiple infection detection (14).
313 Main disadvantages of dPCR are related to the complexity in the experimental setup
314 and the slower turn-around time. However, single molecule measurements decrease
315 inhibition and reagent costs per test, and the high throughput data results in more
316 reliable and accurate sample classification. Coupling dPCR instrument with data-
317 driven multiplexing in diagnostic setups, will facilitate sample identification and
318 generally more robust molecular diagnosis.
319 Although the results in this manuscript present a milestone towards providing a fast,
320 reliable and cost-effective diagnostic solution for multiplexed assays, there are several

321 aspects that need to be further investigated. Firstly, in this paper we evaluate the
322 performance of AMCA method in clinical isolates using pure bacterial cultures. A
323 follow-up study will address the performance of the method directly from clinical
324 samples. Secondly, it is important to identify co-presence of infections for patient
325 treatment, however in this paper we address only one sample with a double infection;
326 a larger study will be required to test the effectiveness of the AMCA in double pathogen
327 identification. Depending on the sample concentration, this might not limit multiplexing
328 capabilities in digital PCR, but it could represent a challenge when qPCR instruments
329 are used.
330 This work suggests that the AMCA approach provides a versatile solution for the rapid
331 detection of AMR genes and delivers reliable, fast and cost-effective results in
332 diagnostic pipelines. This study highlights how important is the integration of artificial
333 intelligence (AI) for diagnosis and how effectively it increases result reliability of state-
334 of-the-art dPCR instruments. Moreover, the AMCA methodology has the potential for
335 further application in point-of-care devices and isothermal chemistries, as a solution
336 to leverage identification accuracy and enable faster detection of several pathogens.
337
338 **Supplemental Material**
339
340 Supplemental material is available at *Clinical Chemistry* online.
341
342 **Nonstandard abbreviations:** qdPCR, real-time digital PCR; dPCR, digital PCR;
343 qPCR, quantitative real-time PCR; LOD, limit of detection, C_q, quantitation cycle; T_m,
344 melting temperature; ICHNT, Imperial College Healthcare NHS Trust; CPO,
345 Carbapenemase-Producing Organisms; AMR, Antimicrobial Resistance: ACA,

346 Amplification Curve Analysis; MCA, Melting Curve Analysis; AMCA, Amplification and
347 Melting Curve Analysis; KNN, k-nearest neighbors.
348
349 **Author Contributions:** *All authors confirmed they have contributed to the intellectual*
350 *content of this paper and have met the following 4 requirements: (a) significant*
351 *contributions to the conception and design, acquisition of data, or analysis and*
352 *interpretation of data; (b) drafting or revising the article for intellectual content; (c) final*
353 *approval of the published article; and (d) agreement to be accountable for all aspects*
354 *of the article thus ensuring that questions related to the accuracy or integrity of any*
355 *part of the article are appropriately investigated and resolved.*
356
357 **Ethical Approval:** All the isolate samples were collected from clinical or screening
358 samples routinely processed by the hospital Microbiology Department at Imperial
359 College Healthcare NHS Trust (Ethics protocol 06/Q0406/20).
360
361 **Data and Code Availability:** All data and code used in this study can be found at:
362 *https://github.com/Luca/pyAMCA_5plex.*
363
364 **Acknowledgments:** This work was supported by NIHR Imperial Biomedical Research
365 Centre [P80763]; [Co-]funded/supported by the Medical Research Council [grant
366 number MR/T005254/1] and EPSRC DTP (EP/ N509486/1 to A.M. and K.M.C.). The
367 research was partially funded by the DHSC Centre for Antimicrobial Optimisation
368 (CAMO) at Imperial College London. The views expressed in this publication are those
369 of the authors and not necessarily those of the Department of Health and Social Care,
370 NHS, or the National Institute for Health Research. HA was sponsored by the Imperial

371 Health Charity; FD is sponsored by the MRC Clinical Academic Research Partnership
372 (CARP) scheme. Initial testing of isolates was performed by North West London
373 Pathology microbiology laboratory. Please note that authors (J.R.M. and A.H.) are
374 affiliated with the National Institute for Health Research Health Protection Research
375 Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance
376 at Imperial College London in partnership with Public Health England (PHE) in
377 collaboration with Imperial Healthcare Partners, University of Cambridge, and
378 University of Warwick. The views expressed in this publication are those of the
379 author(s) and not necessarily those of the NHS, the National Institute for Health
380 Research, the Department of Health and Social Care or Public Health England. A.H.
381 is a National Institute for Health Research (NIHR) senior investigator.
382

383 **References**

384

385 1. Moniri A, Miglietta L, Holmes A, Georgiou P, Rodriguez-Manzano J. High-Level
386 Multiplexing in Digital PCR with Intercalating Dyes by Coupling Real-Time Kinetics
387 and Melting Curve Analysis. Anal Chem. 2020;92:14181–8.
388 2. Tzouvelekis LS, Markogiannakis A, Psichogiou M, Tassios PT, Daikos GL.
389 Carbapenemases in Klebsiella pneumoniae and other Enterobacteriaceae: An
390 evolving crisis of global dimensions. Clin Microbiol Rev. 2012;25:682–707.
391 3. Bush K, Fisher JF. Epidemiological Expansion, Structural Studies, and Clinical
392 Challenges of New β-Lactamases from Gram-Negative Bacteria. Annu Rev Microbiol.
393 Annual Reviews; 2011;65:455–78.
394 4. Otter JA, Doumith M, Davies F, Mookerjee S, Dyakova E, Gilchrist M, et al.
395 Emergence and clonal spread of colistin resistance due to multiple mutational

396 mechanisms in carbapenemase-producing Klebsiella pneumoniae in London. Sci
397 Rep. 2017;7:1–8.
398 5. Rodriguez-Manzano J, Moser N, Malpartida-Cardenas K, Moniri A, Fisarova L,
399 Pennisi I, et al. Rapid Detection of Mobilized Colistin Resistance using a Nucleic Acid
400 Based Lab-on-a-Chip Diagnostic System. Sci Rep. Springer US; 2020;10:1–9.
401 6. Neuner EA, Yeh JY, Hall GS, Sekeres J, Endimiani A, Bonomo RA, et al. Treatment
402 and outcomes in carbapenem-resistant Klebsiella pneumoniae bloodstream
403 infections. Diagn Microbiol Infect Dis. 2011;69:357–62.
404 7. Bleumin D, Cohen MJ, Moranne O, Esnault VLM, Benenson S, Paltiel O, et al.
405 Carbapenem-resistant Klebsiella pneumoniae is associated with poor outcome in
406 hemodialysis patients. J Infect. Elsevier Ltd; 2012;65:318–25.
407 8. Otter JA, Burgess P, Davies F, Mookerjee S, Singleton J, Gilchrist M, et al. Counting
408 the cost of an outbreak of carbapenemase-producing Enterobacteriaceae: an
409 economic evaluation from a hospital perspective. Clin Microbiol Infect. Elsevier Ltd;
410 2017;23:188–96.
411 9. Codjoe F, Donkor E. Carbapenem Resistance: A Review. Med Sci. 2017;6:1.
412 10. Moloney E, Lee KW, Craig D, Allen AJ, Graziadio S, Power M, et al. A PCR-based
413 diagnostic testing strategy to identify carbapenemase-producing Enterobacteriaceae
414 carriers upon admission to UK hospitals: early economic modelling to assess costs
415 and consequences. Diagnostic Progn Res. Diagnostic and Prognostic Research;
416 2019;3:1–9.
417 11. Matsumura Y, Pitout JD. Recent advances in the laboratory detection of
418 carbapenemase-producing Enterobacteriaceae. Expert Rev Mol Diagn. Taylor &
419 Francis; 2016;16:783–94.

420 12. Reta DH, Tessema TS, Ashenef AS, Desta AF, Labisso WL, Gizaw ST, et al.
421 Molecular and Immunological Diagnostic Techniques of Medical Viruses. Int J
422 Microbiol. 2020;2020.
423 13. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models.
424 Lancet. 2019;393:1577–9.
425 14. Moniri A, Miglietta L, Malpartida-Cardenas K, Pennisi I, Cacho-Soblechero M,
426 Moser N, et al. Amplification Curve Analysis: Data-Driven Multiplexing Using Real-
427 Time Digital PCR. Anal Chem. 2020;92:13134–43.
428 15. Moniri A, Rodriguez-Manzano J, Malpartida-Cardenas K, Yu LS, Didelot X,
429 Holmes A, et al. Framework for DNA Quantification and Outlier Detection Using
430 Multidimensional Standard Curves. Anal Chem. 2019;91:7426–34.
431 16. Rodriguez-Manzano J, Moniri A, Malpartida-Cardenas K, Dronavalli J, Davies F,
432 Holmes A, Georgiou P. Simultaneous single-channel multiplexing and quantification
433 of carbapenem-resistant genes using multidimensional standard curves. Anal Chem.
434 2019;91:2013–20.
435 17. Lim YJ, Park HY, Lee JY, Kwak SH, Kim MN, Sung H, et al. Clearance of
436 carbapenemase-producing Enterobacteriaceae (CPE) carriage: a comparative study
437 of NDM-1 and KPC CPE. Clin Microbiol Infect. Elsevier; 2018;24:1104.e5-1104.e8.
438 18. Maurer FP, Castelberg C, Quiblier C, Bloemberg G V., Hombach M. Evaluation of
439 carbapenemase screening and confirmation tests with Enterobacteriaceae and
440 development of a practical diagnostic algorithm. J Clin Microbiol. 2015;53:95–104.
441 19. Athamanolap P, Parekh V, Fraley SI, Agarwal V, Shin DJ, Jacobs MA, et al.
442 Trainable high resolution melt curve machine learning classifier for large-scale reliable
443 genotyping of sequence variants. PLoS One. 2014;9.

444 20. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high
445 throughput. Nucleic Acids Res. 2004;32:1792–7.
446 21. multiple-primer-analyzer @ www.thermofisher.com. Available from:
447 https://www.thermofisher.com/uk/en/home/brands/thermo-scientific/molecular-
448 biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-
449 scientific-web-tools/multiple-primer-analyzer.html
450 22. Corporation F. Digital PCR with the 37k IFC Using Gene-Specific Assays.
451 https://www.fluidigm.com/binaries/content/documents/fluidigm/resources/qdpcr-37k-
452 dpcr-qr-100-6896/qdpcr-37k-dpcr-qr-100-6896
453 23. Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an
454 emphasis to case–control studies. Statistics in medicine. 2007;26:2170–83.
455 24. Huggett JF, Cowen S, Foy CA. Considerations for digital PCR as an accurate
456 molecular diagnostic tool. Clin Chem. 2015;61:79–88.

457 *Table 1. Primer sets developed in this study for the 5-plex assay.*

CPE target	Forward primer sequence (5'—3')	Reverse primer sequence (5'—3')	Amplicon size (bp)	Amplicon T_m (°C)
bla$_{IMP}$	CAGCAGAGYCTTTGCCAGATT	GCCACGYTCCACAAACCAA	203	86.5
bla$_{KPC}$	GGCTCAGGCGCAACTGTAA	GCCCAACTCCTTCAGCAACAA	273	95.5
bla$_{NDM}$	CGCGTGCTGKTGGTCGATA	GGCGAAAGTCAGGCTGTGTTG	240	96
bla$_{OXA-48}$	CGATTTGGGCGTGGTTAAGGAT	GTCGAGCCARAAACTGTCTAC	235	88.5
bla$_{VIM}$	CGAGGYAGAGGGGARCGAGATT	CTSTGCTTCCGGGTAGTGTT	275	94

458

459 *Table 2. Clinical Enterobacteriaceae isolates used in this study.*

Species (MALDI-TOF MS)	Carbapenemase gene	Number of isolates
Citrobacter spp.	*bla*$_{IMP}$	1
	bla$_{KPC}$	2
	bla$_{NDM}$	1
	bla$_{OXA-48}$	10
	bla$_{VIM}$	1
Enterobacter spp.	*bla*$_{IMP}$	20
	bla$_{NDM}$	7
	bla$_{OXA-48}$	2
	bla$_{VIM}$	2
Escherichia spp.	*bla*$_{IMP}$	7
	bla$_{NDM}$	14
	bla$_{NDM}$ & *bla*$_{OXA-48}$	1
	bla$_{OXA-48}$	26
Klebsiella spp.	*bla*$_{IMP}$	15
	bla$_{KPC}$	6
	bla$_{NDM}$	51
	bla$_{OXA-48}$	45
	bla$_{VIM}$	3
Proteus mirabilis	*bla*$_{NDM}$	1
Pseudomonas aeruginosa	*bla*$_{IMP}$	2
	bla$_{VIM}$	2
Serratia marcescens	*bla*$_{KPC}$	1
	bla$_{OXA-48}$	1
Multiple species*	negative	32

460
461 *CPO-negative species: *Acinetobacter baumannii, Citrobacter freundii, Enterobacter spp., Escherichia*
462 *coli, Klebsiella pneumoniaea* and *Pseudomonas aeruginosa.*

463 *Table 3. Classification of clinical isolates based on ML-based MCA method*

Target	N	TP	TN[a]	FP	FN	SEN	SPE	Accuracy (CI)
bla$_{IMP}$	45	45	32	0	0	100.0%	100.0%	100.0% (95.32 to 100.00%)
bla$_{KPC}$	9	8	32	1[b]	0	100.0%	96.97%	97.56% (87.14 to 99.94%)
bla$_{NDM}$	74	54	32	20[c]	0	100.0%	61.54%	81.13% (72.38 to 88.08%)
bla$_{OXA-48}$	84	84	32	0	0	100.0%	100.0%	100.0% (96.87 to 100.00%)
bla$_{VIM}$	8	8	32	0	0	100.0%	100.0%	100.0% (91.19 to 100.00%)
bla$_{OXA-48}$ & *bla*$_{NDM}$	1	1	32	0	0	100.0%	100.0%	100.0% (97.24 to 100.00)
Total	**221**	**200**	**32**	**21**	**0**	**100.0%**	**60.38%**	**91.70% (87.59 to 94.79%)**

464
465 Abbreviations - N: number of samples; TP: True Positive TN: True Negative; FP: False Positive, FN:
466 False Negative, SEN: Sensitivity, SPE: Specificity, CI: Confidence Interval.
467 [a]A total 32 negatives samples are considered across all the groups for sensitivity, specificity and
468 accuracy calculation
469 [b]This isolate was misclassified as *bla*$_{NDM}$ and *bla*$_{KPC}$ double infection
470 [c]These isolates were misclassified as *bla*$_{NDM}$ and *bla*$_{KPC}$ double infections

471 *Table 4. Classification of clinical isolates based on ML-based AMCA method*

Target	N	TP	TN[a]	FP	FN	SEN	SPE	Accuracy (CI)
bla$_{IMP}$	45	45	32	0	0	100.0%	100.0%	100.0% (95.32 to 100.00%)
bla$_{KPC}$	9	9	32	0	0	100.0%	100.0%	100.0% (91.40 to 100.00%)
bla$_{NDM}$	74	73	32	1[b]	0	100.0%	96.97%	99.06% (94.86% to 99.98%)
bla$_{OXA-48}$	84	84	32	0	0	100.0%	100.0%	100.0% (96.87 to 100.00%)
bla$_{VIM}$	8	8	32	0	0	100.0%	100.0%	100.0% (91.19 to 100.00%)
bla$_{OXA-48}$ & *bla*$_{NDM}$	1	1	32	0	0	100.0%	100.0%	100.0% (97.24 to 100.00)
Total	**221**	**220**	**32**	**1**	**0**	**100.0%**	**96.97%**	**99.60% (97.82 to 99.99%)**

472
473 Abbreviations - N: number of samples; TP: True Positive TN: True Negative; FP: False Positive, FN:
474 False Negative, SEN: Sensitivity, SPE: Specificity, CI: Confidence Interval.
475 [a]A total 32 negatives samples are considered across all the groups for sensitivity, specificity and
476 accuracy calculation
477 [b]This isolate was misclassified as *bla*$_{NDM}$ and *bla*$_{KPC}$ double infection

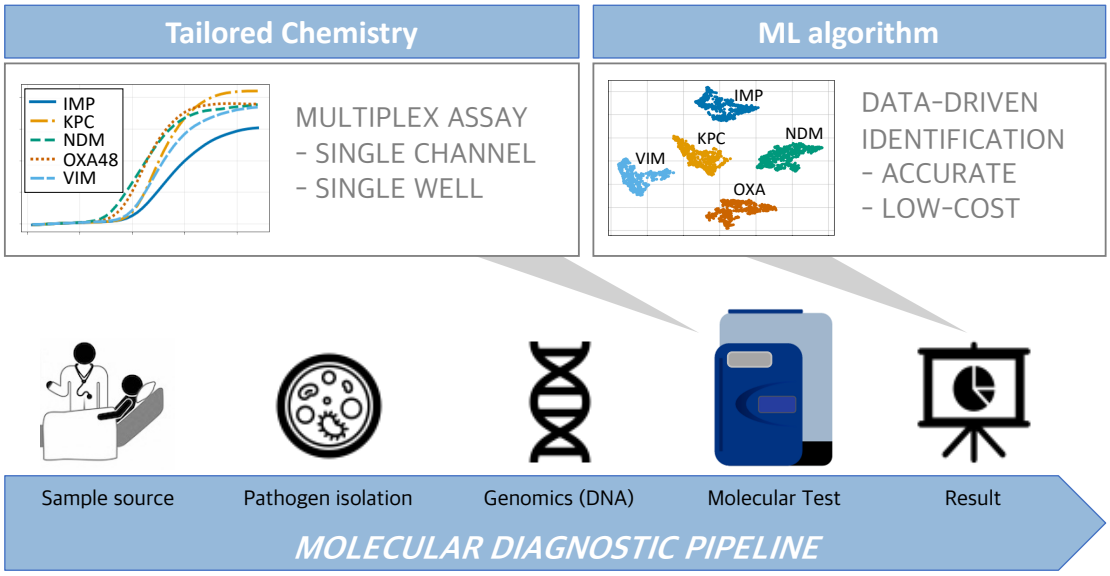

Figure 1. Integration of data-driven approaches to standard diagnostic workflows. The blue arrow indicates the conventional diagnosis pipeline from patient to result. The grey arrow on the left indicates the most suitable genetic test is developed in-silico*, comprising of specialised assays capable of multi target detection in a single reaction in dPCR. The right grey arrow indicates the AMCA methodology is applied for patient result analysis.*

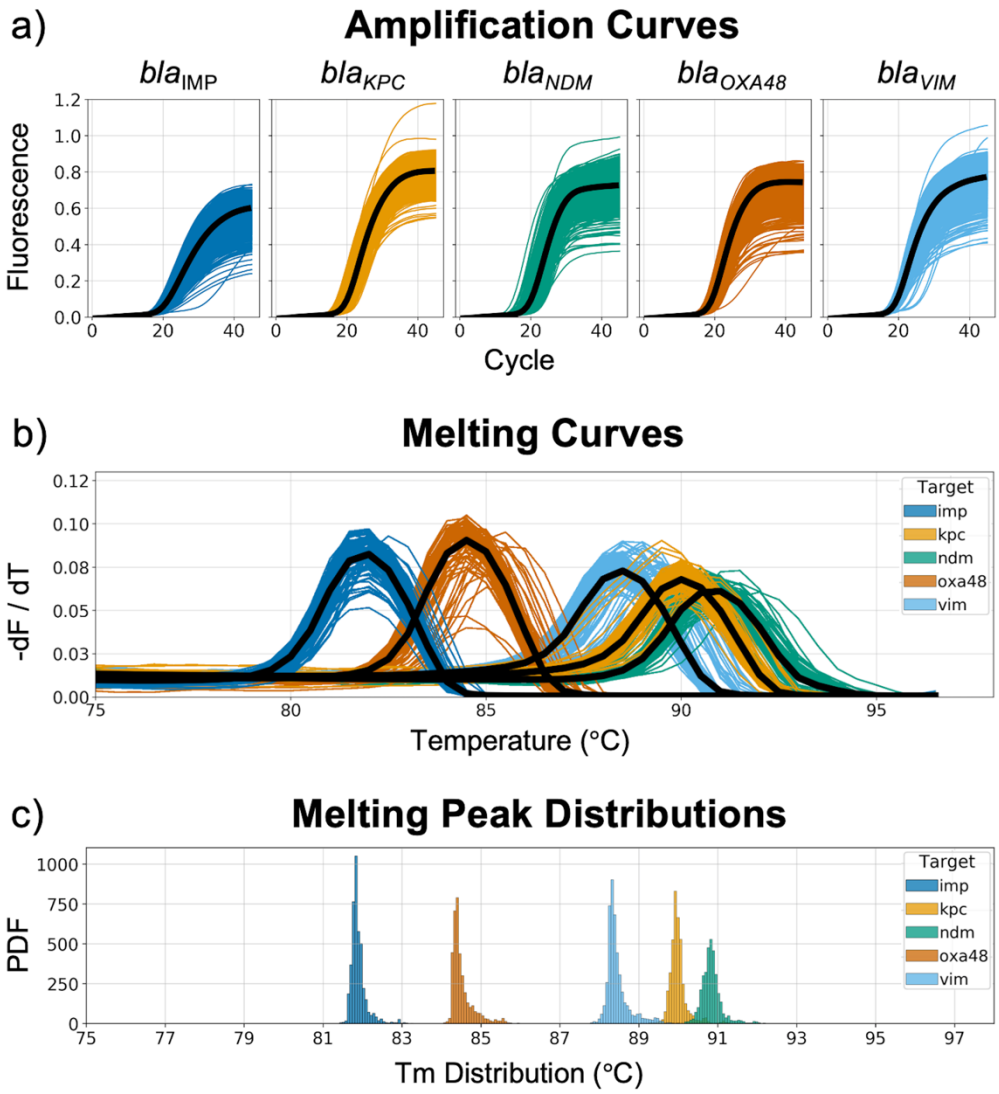

Figure 2. Real-time amplification and melting curves obtained from the dPCR instrument. (a) Raw amplification curves at different concentrations; the black line represents the average trend of the kinetic. (b) Melting curves across the five different CPO; the black line represents the average trend of the thermodynamic. (c) Melting peak (T_m) distribution from the dPCR instrument, showing the probability density function (PDF) for each target.