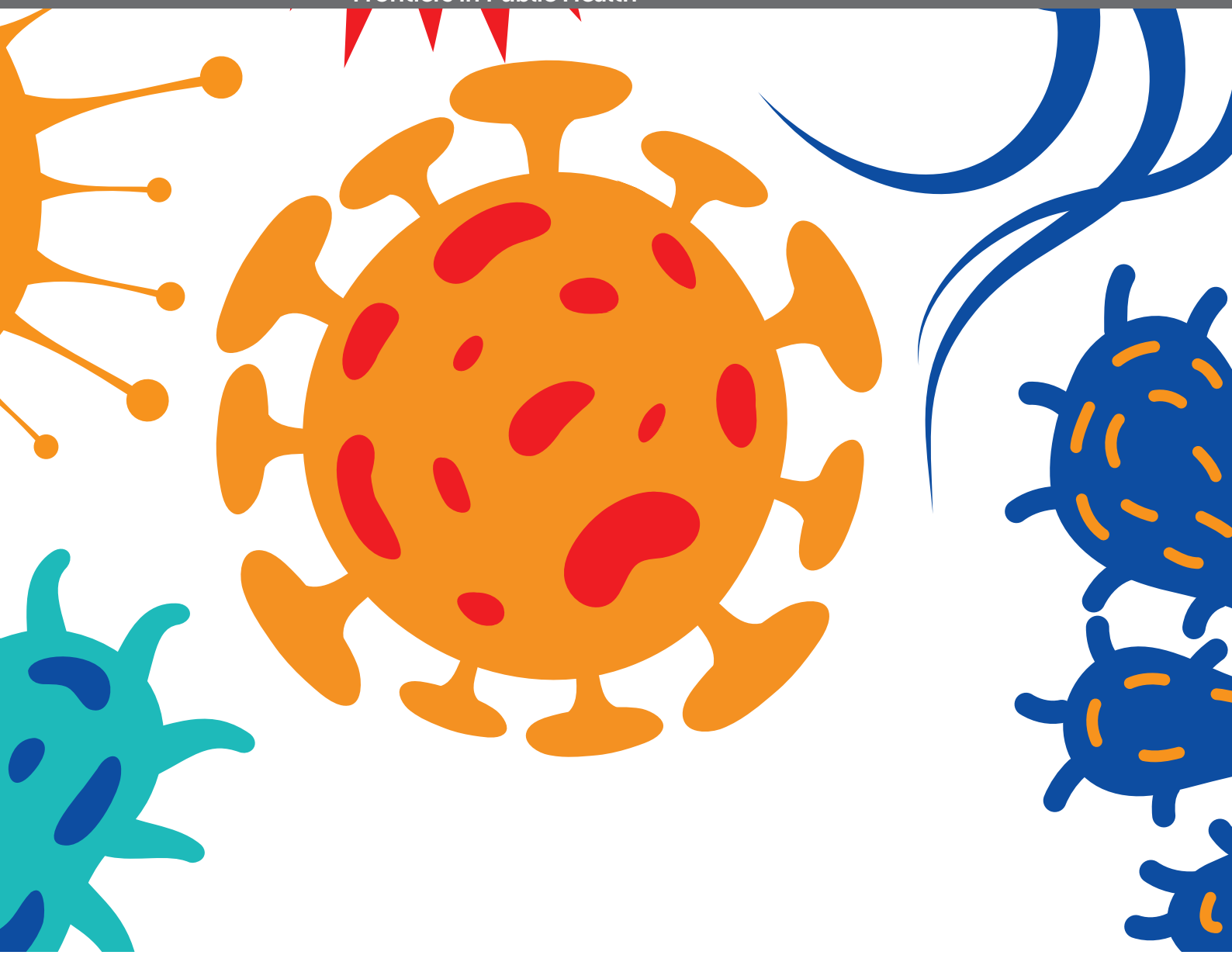
# DIGITALIZATION AND INFECTIOUS DISEASES

EDITED BY: Adrian Egli, Belén Rodriguez-Sanchez and Paul Savelkoul
PUBLISHED IN: Frontiers in Cellular and Infection Microbiology, Frontiers in Digital Health, Frontiers in Medicine and Frontiers in Public Health



frontiers Research Topics

**Frontiers eBook Copyright Statement**

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88974-781-8
DOI 10.3389/978-2-88974-781-8

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

# DIGITALIZATION AND INFECTIOUS DISEASES

Topic Editors:
**Adrian Egli,** University Hospital of Basel, Switzerland
**Belén Rodriguez-Sanchez,** Gregorio Marañón Hospital, Spain
**Paul Savelkoul,** Maastricht University Medical Centre, Netherlands

**Citation:** Egli, A., Rodriguez-Sanchez, B., Savelkoul, P., eds. (2022). Digitalization and Infectious Diseases. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88974-781-8

# Table of Contents

frontiers in Cellular and Infection Microbiology

MINI REVIEW
published: 30 November 2020
doi: 10.3389/fcimb.2020.582028

# Mini Review: Clinical Routine Microbiology in the Era of Automation and Digital Health

*Stefano Leo*[1], *Abdessalam Cherkaoui*[2], *Gesuele Renzi*[2] *and Jacques Schrenzel*[1,2*]

[1] Genomic Research Laboratory, Division of Infectious Diseases, Department of Medicine, Geneva University Hospitals and University of Geneva, Geneva, Switzerland, [2] Bacteriology Laboratory, Division of Laboratory Medicine, Department of Diagnostics, Geneva University Hospitals, Geneva, Switzerland

OPEN ACCESS

**Edited by:**
Belén Rodriguez-Sanchez, Gregorio Marañón Hospital, Spain

**Reviewed by:**
Charles William Stratton, Vanderbilt University Medical Center, United States
Elena De Carolis, Catholic University of the Sacred Heart, Italy

***Correspondence:**
Jacques Schrenzel
jacques.schrenzel@hcuge.ch

**Specialty section:**
This article was submitted to Clinical Microbiology, a section of the journal Frontiers in Cellular and Infection Microbiology

**Received:** 10 July 2020
**Accepted:** 20 October 2020
**Published:** 30 November 2020

**Citation:**
Leo S, Cherkaoui A, Renzi G and Schrenzel J (2020) Mini Review: Clinical Routine Microbiology in the Era of Automation and Digital Health. Front. Cell. Infect. Microbiol. 10:582028. doi: 10.3389/fcimb.2020.582028

Clinical microbiology laboratories are the first line to combat and handle infectious diseases and antibiotic resistance, including newly emerging ones. Although most clinical laboratories still rely on conventional methods, a cascade of technological changes, driven by digital imaging and high-throughput sequencing, will revolutionize the management of clinical diagnostics for direct detection of bacteria and swift antimicrobial susceptibility testing. Importantly, such technological advancements occur in the golden age of machine learning where computers are no longer acting passively in data mining, but once trained, can also help physicians in making decisions for diagnostics and optimal treatment administration. The further potential of physically integrating new technologies in an automation chain, combined to machine-learning-based software for data analyses, is seducing and would indeed lead to a faster management in infectious diseases. However, if, from one side, technological advancement would achieve a better performance than conventional methods, on the other side, this evolution challenges clinicians in terms of data interpretation and impacts the entire hospital personnel organization and management. In this mini review, we discuss such technological achievements offering practical examples of their operability but also their limitations and potential issues that their implementation could rise in clinical microbiology laboratories.

**Keywords: clinical microbiology, machine learning, laboratory automation, diagnostics, next-generation sequencing**

## INTRODUCTION

Fully automated diagnostics pipeline is a seducing idea and first automated microbiology laboratories have started to be implemented world-wide (Vandenberg et al., 2018; Vandenberg et al., 2020). In parallel, machine learning (ML), a branch of artificial intelligence, has gained a foothold in many fields of clinical medicine (Topol, 2019). We actually have ML-driven tools that

can make diagnosis, help clinicians in decision-making challenges (Peiffer-Smadja et al., 2020), such as the choice for a given treatment, and even empower the patients themselves to manage their healthcare (Topol, 2019). The innovative aspect of ML is that it is not a ruled-based system; ML algorithms can learn from input data and automatically make predictions or decisions.

With next-generation sequencing (NGS) techniques, we can gain information about pathogens analyzing millions of small fragments coming from their genomes and even gain insights on microbiota composition, including not-yet cultured or uncultivable organisms.

Can automation, together with new technologies, make a difference from conventional clinical microbiology tests that often require a significant amount of manual work?

What impact will such advancements have in clinical routine in terms of sample-to-result timing, taking into account that it usually takes between 24 and 48 h to obtain results in current routine laboratories (Ruppé et al., 2016)? What will such new technologies imply in terms of resources and management? Lastly, can we understand and interpret multimodal large-volume data resulting from these new technologies?

In this mini review, we will discuss these questions leveraging the benefits of technological advancements over routine diagnostics but also considering the limitations and problems by implementing them in healthcare facilities.

## FULL AUTOMATION IN CLINICAL MICROBIOLOGY LABORATORIES

In a clinical microbiology routine laboratory, sample processing varies mostly because of the nature of the specimens (blood, urine, etc.) but also because of the diversity of pathogens that can require specific media and growth conditions. Besides pathogen identification, clinical microbiology laboratories are also in charge of providing information about the antibiotic susceptibility of pathogens to help selecting the most appropriate pharmacological regimen. Antibiotic susceptibly tests (ASTs) can be performed with different approaches (agar disk diffusion, agar gradient diffusion or broth microdilution) and can measure the minimum inhibitory concentration (MIC) of an antibiotic, that is the lowest concentration of the drug at which there is no visible growth.

To date there are only two commercially available instruments, the Copan's WASPLab™ (WASPLab™) and the Becton Dickinson's Kiestra TLA (Kiestra TLA), which propose automated culture-based tests including specimen streaking, slide preparation, transfer of inoculated media between instruments and automated incubators (Dauwalder et al., 2016; Bailey et al., 2019).

The WASPLab™ and Kiestra TLA are versatile technologies which can incorporate or can be combined with other diagnostic systems such as MALDI-TOF (Cherkaoui et al., 2011; Mutters et al., 2014), a key technique in modern medical microbiology to identify bacteria and fungi (Cherkaoui et al., 2010; Kaleta et al., 2011; Clark et al., 2013; Patel, 2019; Cherkaoui et al., 2020a). For example, the Kiestra TLA combined with MALDI-TOF has been shown to shorten the incubation time required to identify microbial pathogens (Mutters et al., 2014). Unlike Kiestra TLA, WASPLab™ offers an automated solution for antimicrobial disc diffusion susceptibility testing with equal or better accuracy than other available phenotypic methods (Cherkaoui et al., 2020b).

Overall, the two systems reduce the number of manual pre-analytic, analytic and post-analytic steps that are typically performed in a non-automated laboratory (Dauwalder et al., 2016). The implementation of the WASPLab™ or of the Kiestra TLA systems in clinical settings improved sample processing steps and reduced sample-to-result timing (Barake et al., 2017; Cherkaoui et al., 2019a; Cherkaoui et al., 2020c).

Since 2018, the Copan's WASPLab™ technology has been implemented at the Geneva University Hospitals (Hôpitaux Universitaires de Genève—HUG) (Cherkaoui et al., 2020c), where it has proven offering rapid detection of vancomycin-resistant enterococci with automated incubation and digital-image based analysis system (Cherkaoui et al., 2019b) and more generally, a substantial shortening of turn-around times (Cherkaoui et al., 2019a; Cherkaoui et al., 2020a).

Full automation of diagnostic procedures can generate further advantages (Dauwalder et al., 2016; Cherkaoui et al., 2020c).

Firstly, automation increases the capability of sample processing with a better documentation and traceability. Secondly, there is a better control of the costs (e.g. reagents, medium, etc.) with reduced turn-around times thus resulting in a faster diagnosis. Thirdly, full automation permits extending the opening hours of the laboratory with a huge benefit for patient care.

Hopefully full automation will also incorporate molecular diagnostic capabilities, starting with DNA extraction, another procedure that is multi-step and requires experienced technical personnel.

Nowadays, there are plenty of DNA processing machines ranging from low to medium- and high-throughput, but not yet included in Kiestra TLA nor in WASPLab™ systems. In particular, we can distinguish two main types of instruments among commercially available ones: one that combines DNA extraction with the amplification, and the other one where extraction and amplification are performed separately (Ali et al., 2017; Shin, 2018). A technology based on automated nucleic acids (NA) analyses would be advantageous in those situations where NA-based testing is demanded on a large scale, like SARS-CoV-2 pandemic, and offering additional consolidation.

## NEXT-GENERATION SEQUENCING TECHNOLOGIES

NGS has represented a further milestone in clinical microbiology. Today we have four main sequencing technologies, Illumina, Ion Torrent, Pacific Biosciences (PacBio) and Oxford Nanopore (**Figure 1**), which are based on a different chemistry for the sequencing and that provide different outputs in terms of number and length of the sequencing reads. Currently, Illumina

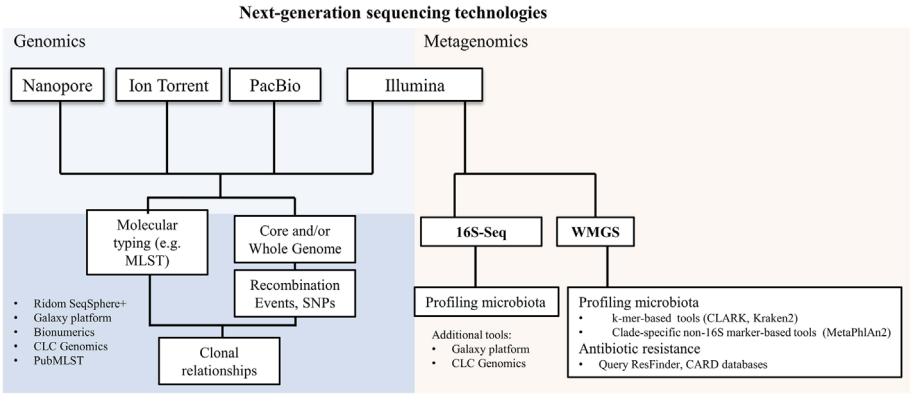



**FIGURE 1** | Next-generation sequencing technologies and their applications in microbiology. A non-exhaustive list of bioinformatics tools used for genomics and metagenomics analyses is reported. SNPs, single nucleotide polymorphisms; 16S-Seq, 16S-sequencing; WMGS, whole metagenome shotgun sequencing.

short-read sequencing is the most used technology for both genomics and metagenomics, due to its sequencing depth and therefore accuracy (**Figure 1**). However, the speed of sequencing of Oxford Nanopore, combined with its ability to sequence long reads, makes it also very compelling for some diagnostic procedures (Grädel et al., 2020).

Parallel to the sequencing technological advancements, there has been an explosion of bioinformatics tools that are capable to analyze and structure the information from sequencing data.

While some of these tools, such as Galaxy platform (Giardine et al., 2005), Ridom SeqSphere+ (Ridom GmbH), CLC Genomics Workbench 20.0 (QIAGEN) and BioNumerics (Applied Maths NV - bioMérieux) display graphical user interfaces, there are many others which require coding skills for their proper and powerful usage. Most codes are publicly shared in open repositories such as GitHub and Bitbucket.

We can today apply NGS to study the core and/or whole genome (Genomics; **Figure 1**) to infer any kind of molecular typing from MLST to vaccine antigens (Pérez-Losada et al., 2018; Muzzi et al., 2019; Leo et al., 2020) and even study clonal relationships by investigating single nucleotide polymorphisms (SNPs) or genomic recombination events (Didelot and Wilson, 2015; Donner et al., 2020; Olearo et al., 2020; Pham et al., 2020; Scherrer et al., 2020).

A further important application of NGS, called metagenomics, is to profile microbiota. Metagenomics has linked microbiota species composition to a broad range of infectious diseases (Forbes et al., 2018; Egli et al., 2020), including complex nosocomial infections as ventilator-associated pneumonia (Emonet et al., 2019), suspected infectious endocarditis (Choutko et al., 2019; Kolb et al., 2019), or challenging deep-seated infections (Lazarevic et al., 2018; Foulex et al., 2019).

Metagenomics consists of two largely used experimental methods: amplicon-based (targeted metagenomics, also called metataxonomics) and whole metagenome shotgun sequencing (WMGS) (**Figure 1**). Targeted metagenomics is based on the amplification, followed by sequencing, of hypervariable regions in a target gene present in all species of the same kingdom. The gene encoding for 16S ribosomal RNA is the most used to generate taxonomic profiles. Bacterial detection by 16S-sequencing can be limited to taxonomic levels higher than the species level in some cases; besides it excludes viruses and fungi from the analyses.

Sequencing reads generated by WMGS are queried against large databases and eventually assigned to a given species not only from bacteria but also from other organisms, including Archaea, DNA viruses and eukaryotic microbes. The relative abundance of species is used to quantify a species with respect to the amount of sequencing reads.

Two main approaches are used for species identification in metagenomic sequencing datasets: k-mers- and clade-specific-marker-based. Beyond purely technical aspects, the main difference between the two methods is that k-mers-based tools, like CLARK (Ounit et al., 2015) and Kraken2 (Wood and Salzberg, 2014), can be used for large customized genome databases, while marker-based approaches, like MetaPhlAn2 (Truong et al., 2015), rely on the querying of reads against a more limited gene sequence dataset. The result is that we can detect a wider range of species with k-mers-based tools than with a marker-based approach (Leo et al., 2017). A further application of WMGS is to search for genetic antibiotic resistance by querying antibiotic resistance gene databases, like ResFinder (Zankari et al., 2012) and the Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al., 2013).

Metagenomics is an appealing tool for the diagnosis of infectious diseases as it has shown to be functionally equivalent to culture techniques (Leo et al., 2017), but it can detect pathogens when they are missed by current laboratory methods (Xu et al., 2011; Mokili et al., 2013; Wan et al., 2013); it could also constitute a promising tool to be integrated in infection control and clinical epidemiology (Greninger et al., 2015).

NGS and metagenomics have not yet been automatized and the utilization of ML has been applied to different aspects, as

inferring antibiotic resistance, predicting diagnosis and recurrent infection (Peiffer-Smadja et al., 2020).

## ARTIFICIAL INTELLIGENCE IN AUTOMATED CLINICAL MICROBIOLOGY DIAGNOSTICS

Together with automation and NGS, artificial intelligence could also contribute to a better management of infectious diseases in helping clinicians to collect and elaborate information from clinical tests.

Computer vision that is the ability of a computer to process a digital image and identify objects represents one of the most popular examples of how artificial intelligence works. In clinical microbiology field, computer vision can be useful to improve the identification of pathogens with all those tasks that are manual and require a certain expertise like the interpretation of Gram stains (Dauwalder et al., 2016).

In fact Gram stain is an essential test which provides initial information on the presence and type of bacteria and helps in opting for a first prompt antibiotic regimen (Barenfanger et al., 2008). Smith and Kang et al. (Smith et al., 2018) realized a system where both slide imaging and Gram stain analyses interpretation were automated. They used a ML algorithm that can analyze digital images and recognize most common pathogens of bloodstream infections based on their morphologies. Their automated ML system reached an accuracy of 92.5% compared to manual classification. Similar results were obtained by adopting ML approaches to automate antimicrobial susceptibility testing and the definition of antimicrobial minimal inhibitory concentrations on the five most common Gram-negative pathogens *Escherichia coli*, *Enterobacter cloacae*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, and *Acinetobacter baumannii* (Smith et al., 2017).

Computer vision can ideally be applied to any type of morphologic/phenotypic test, including parasitological ones. For example, ML was applied to identify parasitic protozoa from fecal matter (Mathison et al., 2020) and malaria parasites (Florin et al., 2018).

Beyond facilitating the automation of certain tasks, ML can be of help in saving time and expenses in clinical laboratories. Burton et al. (2019) applied ML algorithms to predict whether urine samples required further testing by considering not only biological matter present in the sample (counts of white, red blood and epithelial cells) but also other factors like the pregnancy status or the age of the patient.

A recent work (Mueller et al., 2020) describes how a computer tool could analyze and validate the amplification curves generated from reverse transcription polymerase chain reaction (RT-PCR) developed for SARS-CoV-2 testing. In fact, the validation of these laboratory tests can become a laborious task for clinical personnel especially when they are performed on large scale. The consequence is to slow down the delivery of the test outcome to the patient. The algorithm developed by Mueller et al. (2020) can automatically validate SARS-CoV-2 RT-PCR tests and retain those that need particular attention.

In this perspective, such computer-based tools would help focusing on the cases that need further microbiological investigation.

## IMPLEMENTING NEW TECHNOLOGIES IN REAL-WORLD SETTINGS: CONSIDERATIONS AND LIMITATIONS

The implementation of new technologies, like automation, ML and NGS, brings several issues. Automation of a clinical microbiology laboratory is challenging until it can reach all the steps, like opening all routinely used sample containers, relying on validated incubation times and standardized antibiotic susceptibility testing (Dauwalder et al., 2016; Cherkaoui et al., 2019a; Cherkaoui et al., 2020a; Vandenberg et al., 2020).

Standardization and validation of the pre-analytical, analytical and post-analytical procedures are needed before the automated system is fully applicable to routine analyses. In this respect, tasks of the automated pipeline could be segmented and sequentially validated allowing also a better management of personnel training and implementation of instruments in the hospital routine daily life (Cherkaoui et al., 2020c). Importantly an appropriate IT system should be put in place to ensure a correct information exchange with the automated system, e.g. for the protocol of the microbiological tests/tasks to perform (Cherkaoui et al., 2020c).

Biosafety is also an important aspect that should be carefully considered when implementing a new system to appropriately handle clinical samples with biological hazard, in order to prevent accidental infections among laboratory personnel or laboratory contaminations.

ML-driven technologies are "black boxes", meaning that the processes leading from the input to the output are unknown to the user. Therefore, although ML represents a promising tool especially in coping with large-volume complex data, the understanding of its functioning might be hard for microbiologists and clinicians who must inspect and validate the results. Furthermore, ML-driven technologies should be examined in clinical trials in order to be safely and officially incorporated in laboratory-certified operations. Thus, whether ML approaches bring an added value to diagnostics remains to be clarified, once routine implementation can be achieved and potential benefits measured.

NGS and metagenomics are neither fully standardized, nor streamlined in a way that they can smoothly integrate a routine microbiology laboratory. Some efforts to converge towards national/international validated procedures have been undertaken (Ruppé et al., 2017; Ruppé and Schrenzel, 2018; Ruppé and Schrenzel, 2019; Charretier et al., 2020). Moreover, given the large volume of sequencing data, metagenomics can demand a lot of computing resources and can be time-consuming. NGS can detect species in terms of "relative abundance" to which we should

find a meaningful corresponding parameter to allow comparison with culture data.

Automated systems and NGS require the availability of suitable host facilities, trained personnel and adequate informatics infrastructure for data computation, analysis, interpretation and storage. In the absence of such factors, small hospitals are excluded from these technological advancements. Therefore, a reorganization of diagnostics laboratory networking is warranted. Although different models of automated clinical microbiology laboratories are currently implemented (Vandenberg et al., 2020), they are all characterized by a central facility with one or more satellite laboratories. While the central facility should incorporate all the current key technologies, including automatized system and NGS, satellite laboratories serve as platforms for rapid response tests (Vandenberg et al., 2020).

Particular attention should be put at data communication and sharing. We can imagine that these exchanges develop at three different levels (**Figure 2**): 1) between personnel (clinicians, laboratory operators) belonging to the same hospital facility; 2) between personnel from satellite and central facilities of the same hospital corporation; and 3) between different hospitals.

For level 1), video platforms, like Zoom or Skype, provided that they respect the required medical confidentiality, might be considered for rapid clinical consultations and thus valuable instruments to keep communication during unusual situation such as the COVID-19 pandemic.

Irrespective of the type of relationships between facilities, digitalization should be accompanied with appropriate data reporting and rigorous regulation of patient data sharing.

Electronic health record (EHR) is the systematic collection of patient information in digital machine-readable format and represents a solution to data communication and interoperability between the disparate hospitals, on condition that consistent ontology definitions are used. The FAIR (Findability, Accessibility, Interoperability and Reusability) initiative principles (Wilkinson et al., 2016) should be considered to generate formal diagnostic concepts and to define standard diagnostic definitions used in EHRs. A constant curation and revision of ontologies should then be ensured especially when new technologies are introduced in routine analyses. This is the case of genomics, where information are very often not structured in a machine-readable format where new technical terms (Mascia et al., 2018) and new types of data representation are introduced. Therefore, the constitution of a data report for genomic data which is largely understood and accepted by the clinicians should be evaluated (Crisan et al., 2018).

Exchange of clinical data between infrastructures implies that patient privacy should be guaranteed at any operation level and an *ad hoc* security system should be used. Privacy-protecting technologies like homomorphic encryption and secure multiparty computations could ensure a protected environment where to store or locally analyze data, that is without the need to electronically transfer them to another informatics environment (Grishin et al., 2019). Implementation of secure computation, based on cryptographic protocol that covers the features of patients, has also been proposed for the analyses of microbiome (Wagner et al., 2016).

Initiatives like the Global Alliance for Genomics and Health (https://www.ga4gh.org/) and the European Union General Data Protection Regulation (https://eugdpr.org/), aim to harmonize legislation concerning the treatment and the protection of clinical genomic data. In Switzerland, the BioMedIT project (https://sphn.ch/network/projects/biomedit/) was established for a secure national coordination and transmission of clinical information among biomedical infrastructures.

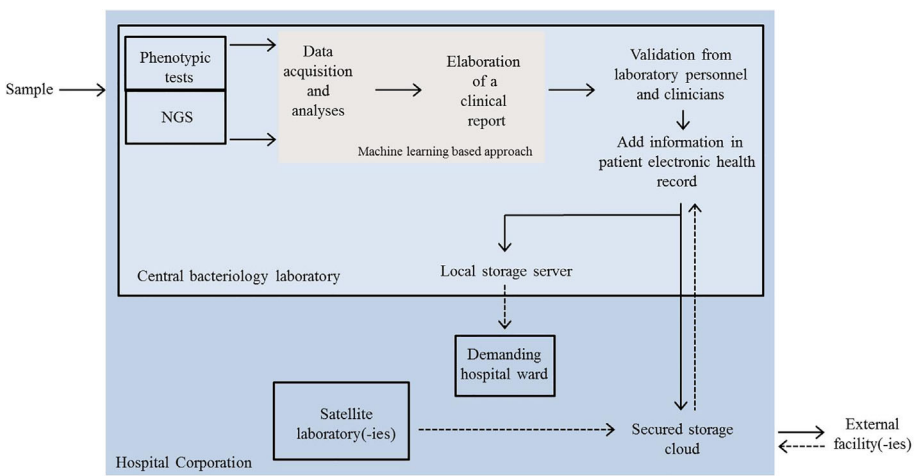



**FIGURE 2** | Schematic representation of a possible future scenario in the dynamics of automated clinical microbiology laboratory networking. Clinical samples are analysed by automated phenotypic tests or by NGS at the central bacteriology laboratory. Data acquisition, mining and elaboration of a first clinical report are performed by a machine learning approach. The final report is evaluated by technical and clinician experts and resulting information added to an electronic health record (EHR). EHR is then shared either internally (local server) or sent outside. Satellite laboratories and external facilities can also send the outcomes of rapid tests or other analyses to the central facility *via* a secured cloud and newly acquired information can be integrated in EHRs. NGS, next-generation sequencing.

## CONCLUSIONS

New technological advancements are going to change the appearance of clinical microbiology routine laboratories with data increasing in volume and complexity. Yet, their implementation in real clinical settings should still prove an improvement in making processes faster and cleaner than conventional workflows. Explainability and interpretability of ML-based tools are rarely addressed and independent validations should be carried out. A re-arrangement of local and regional diagnostics facilities is demanded to better cover the needs of management of automated laboratories.

## AUTHOR CONTRIBUTIONS

SL, AC, GR, and JS conceptualized and wrote the manuscript. All authors contributed to the article and approved the submitted version.

## ACKNOWLEDGMENTS

We would like to thank the reviewers for the important suggestions and comments to improve the manuscript. We would like to apologize to colleagues whose work could not be cited due to space constraints.

## REFERENCES

Ali, N., Rampazzo, R. C. P., Costa, A. D. T., and Krieger, M. A. (2017). Current nucleic acid extraction methods and their implications to point-of-care diagnostics. *BioMed. Res. Int.* 2017, 9306564–9306564. doi: 10.1155/2017/9306564

Bailey, A. L., Ledeboer, N., and Burnham, C.-A. D. (2019). Clinical microbiology is growing up: the total laboratory automation revolution. *Clin. Chem.* 65 (5), 634–643. doi: 10.1373/clinchem.2017.274522

Barake, S. S., Emrick, A., Tabak, Y., Jasen, A., Vankeepuram, L., Sellers, D., et al. (2017). Impact of automation process on microbiological laboratory efficiency. *Open Forum Infect. Dis.* 4 (Suppl 1), S593–S593. doi: 10.1093/ofid/ofx163.1555

Barenfanger, J., Graham, D. R., Kolluri, L., Sangwan, G., Lawhorn, J., Drake, C. A., et al. (2008). Decreased mortality associated with prompt Gram staining of blood cultures. *Am. J. Clin. Pathol.* 130 (6), 870–876. doi: 10.1309/AJCPVMDQU2ZJDPBL

Burton, R. J., Albur, M., Eberl, M., and Cuff, S. M. (2019). Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. *BMC Med. Inf. Decis. Mak.* 19 (1), 171. doi: 10.1186/s12911-019-0878-9

Charretier, Y., Lazarevic, V., Schrenzel, J., and Ruppé, E. (2020). Messages from the Fourth International Conference on Clinical Metagenomics. *Microbes Infect.* doi: 10.1016/j.micinf.2020.07.007

Cherkaoui, A., Hibbs, J., Emonet, S., Tangomo, M., Girard, M., Francois, P., et al. (2010). Comparison of two matrix-assisted laser desorption ionization-time of flight mass spectrometry methods with conventional phenotypic identification for routine identification of bacteria to the species level. *J. Clin. Microbiol.* 48 (4), 1169–1175. doi: 10.1128/JCM.01881-09

Cherkaoui, A., Emonet, S., Fernandez, J., Schorderet, D., and Schrenzel, J. (2011). Evaluation of matrix-assisted laser desorption ionization-time of flight mass spectrometry for rapid identification of beta-hemolytic streptococci. *J. Clin. Microbiol.* 49 (8), 3004–3005. doi: 10.1128/JCM.00240-11

Cherkaoui, A., Renzi, G., Vuilleumier, N., and Schrenzel, J. (2019a). Copan WASPLab automation significantly reduces incubation times and allows earlier culture readings. *Clin. Microbiol. Infect.* 25 (11), 1430.e1435–1430.e1412. doi: 10.1016/j.cmi.2019.04.001

Cherkaoui, A., Renzi, G., Charretier, Y., Blanc, D. S., Vuilleumier, N., and Schrenzel, J. (2019b). Automated incubation and digital image analysis ofchromogenic media using Copan WASPLab enables rapid detection of vancomycin-resistant Enterococcus. *Front. Cell. Infect. Microbiol.* 9, 379. doi: 10.3389/fcimb.2019.00379

Cherkaoui, A., Renzi, G., Azam, N., Schorderet, D., Vuilleumier, N., and Schrenzel, J. (2020a). Rapid identification by MALDI-TOF/MS and antimicrobial disk diffusion susceptibility testing for positive blood cultures after a short incubation on the WASPLab. *Eur. J. Clin. Microbiol. Infect. Dis.* 39 (6), 1063–1070. doi: 10.1007/s10096-020-03817-8

Cherkaoui, A., Renzi, G., Fischer, A., Azam, N., Schorderet, D., Vuilleumier, N., et al. (2020b). Comparison of the Copan WASPLab incorporating the BioRad expert system against the SIRscan 2000 automatic for routine antimicrobial disc diffusion susceptibility testing. *Clin. Microbiol. Infect.* 26 (5), 619–625. doi: 10.1016/j.cmi.2019.11.008

Cherkaoui, A., Renzi, G., Viollet, A., Fleischmann, M., Metral-Boffod, L., Dominguez-Amado, D., et al. (2020c). Implementation of the WASPLab™ and first year achievements within a university hospital. *Eur. J. Clin. Microbiol. Infect. Dis* 39, 1527–1534. doi: 10.1007/s10096-020-03872-1

Choutko, V., Lazarevic, V., Gaïa, N., Girard, M., Renzi, G., Leo, S., et al. (2019). Rare case of community-acquired endocarditis caused by Neisseria meningitidis assessed by clinical metagenomics. *Front. Cardiovasc. Med.* 6, 112–112. doi: 10.3389/fcvm.2019.00112

Clark, A. E., Kaleta, E. J., Arora, A., and Wolk, D. M. (2013). Matrix-assisted laser desorption ionization-time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology. *Clin. Microbiol. Rev.* 26 (3), 547–603. doi: 10.1128/CMR.00072-12

Crisan, A., McKee, G., Munzner, T., and Gardy, J. L. (2018). Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. *PeerJ* 6, e4218–e4218. doi: 10.7717/peerj.4218

Dauwalder, O., Landrieve, L., Laurent, F., de Montclos, M., Vandenesch, F., and Lina, G. (2016). Does bacteriology laboratory automation reduce time to results and increase quality management? *Clin. Microbiol. Infect.* 22 (3), 236–243. doi: 10.1016/j.cmi.2015.10.037

Didelot, X., and Wilson, D. J. (2015). ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PloS Comput. Biol.* 11 (2), e1004041. doi: 10.1371/journal.pcbi.1004041

Donner, V., Buzzi, M., Lazarevic, V., Gaïa, N., Girard, M., Renzi, F., et al. (2020). Septic shock caused by Capnocytophaga canis after a cat scratch. *Eur. J. Clin. Microbiol. Infect. Dis.* 39 (10), 1993–1995. doi: 10.1007/s10096-020-03922-8

Egli, A., Koch, D., Danuser, J., Hendriksen, R. S., Driesen, S., Schmid, D. C., et al. (2020). Symposium report: One Health meets sequencing. *Microbes Infect.* 22 (1), 1–7. doi: 10.1016/j.micinf.2019.07.004

Emonet, S., Lazarevic, V., Leemann Refondini, C., Gaïa, N., Leo, S., Girard, M., et al. (2019). Identification of respiratory microbiota markers in ventilator-associated pneumonia. *Intensive Care Med.* 45 (8), 1082–1092. doi: 10.1007/s00134-019-05660-8

Florin, L., Maelegheer, K., Muyldermans, A., Van Esbroeck, M., Nulens, E., and Emmerechts, J. (2018). Evaluation of the CellaVision DM96 advanced RBC application for screening and follow-up of malaria infection. *Diagn. Microbiol. Infect. Dis.* 90 (4), 253–256. doi: 10.1016/j.diagmicrobio.2017.12.002

Forbes, J. D., Knox, N. C., Peterson, C.-L., and Reimer, A. R. (2018). Highlighting clinical metagenomics for enhanced diagnostic decision-making: A Step Towards Wider Implementation. *Comput. Struct. Biotechnol. J.* 16, 108–120. doi: 10.1016/j.csbj.2018.02.006

Foulex, A., Coen, M., Cherkaoui, A., Lazarevic, V., Gaïa, N., Leo, S., et al. (2019). Listeria monocytogenes infectious periaortitis: a case report from the infectious disease standpoint. *BMC Infect. Dis.* 19 (1), 326–326. doi: 10.1186/s12879-019-3953-z

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 15 (10), 1451–1455. doi: 10.1101/gr.4086505

Grädel, C., Terrazos Miani, M. A., Baumann, C., Barbani, M. T., Neuenschwander, S., Leib, S. L., et al. (2020). Whole-genome sequencing of human Enteroviruses from clinical samples by Nanopore direct RNA sequencing. *Viruses* 12 (8), 841. doi: 10.3390/v12080841

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., et al. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.* 7, 99. doi: 10.1186/s13073-015-0220-9

Grishin, D., Obbad, K., and Church, G. M. (2019). Data privacy in the age of personal genomics. *Nat. Biotechnol.* 37 (10), 1115–1117. doi: 10.1038/s41587-019-0271-3

Kaleta, E. J., Clark, A. E., Cherkaoui, A., Wysocki, V. H., Ingram, E. L., Schrenzel, J., et al. (2011). Comparative analysis of PCR–electrospray ionization/mass spectrometry (MS) and MALDI-TOF/MS for the identification of bacteria and yeast from positive blood culture bottles. *Clin. Chem.* 57 (7), 1057–1067. doi: 10.1373/clinchem.2011.161968

Kolb, M., Lazarevic, V., Emonet, S., Calmy, A., Girard, M., Gaïa, N., et al. (2019). Next-generation sequencing for the diagnosis of challenging culture-negative endocarditis. *Front. Med.* 6, 203–203. doi: 10.3389/fmed.2019.00203

Lazarevic, V., Gaïa, N., Girard, M., Leo, S., Cherkaoui, A., Renzi, G., et al. (2018). When bacterial culture fails, metagenomics can help: a case of chronic hepatic Brucelloma assessed by next-generation sequencing. *Front. Microbiol.* 9 (1566). doi: 10.3389/fmicb.2018.01566

Leo, S., Gaïa, N., Ruppé, E., Emonet, S., Girard, M., Lazarevic, V., et al. (2017). Detection of bacterial pathogens from broncho-alveolar lavage by next-generation sequencing. *Int. J. Mol. Sci.* 18. doi: 10.3390/ijms18092011

Leo, S., Lazarevic, V., Girard, M., Getaz-Jimenez Velasco, G. C., Gaïa, N., Renzi, G., et al. (2020). Strain coverage of Bexsero vaccine assessed by whole-genome sequencing over a cohort of invasive meningococci of serogroups B and W isolated in Switzerland. *Vaccine* 38 (33), 5324–5331. doi: 10.1016/j.vaccine.2020.05.071

Mascia, C., Uva, P., Leo, S., and Zanetti, G. (2018). OpenEHR modeling for genomics in clinical practice. *Int. J. Med. Inf.* 120, 147–156. doi: 10.1016/j.ijmedinf.2018.10.007

Mathison, B. A., Kohan, J. L., Walker, J. F., Smith, R. B., Ardon, O., and Couturier, M. R. (2020). Detection of intestinal Protozoa in Trichrome-stained stool specimens by use of a deep convolutional neural network. *J. Clin. Microbiol.* 58 (6), e02053–e02019. doi: 10.1128/JCM.02053-19

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., et al. (2013). : The Comprehensive Antibiotic Resistance Database. *Antimicrob. Agents Chemother.* 57 (7), 3348. doi: 10.1128/AAC.00419-13

Mokili, J. L., Dutilh, B. E., Lim, Y. W., Schneider, B. S., Taylor, T., Haynes, M. R., et al. (2013). Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. *PloS One* 8 (3), e58404. doi: 10.1371/journal.pone.0058404

Mueller, L., Scherz, V., Greub, G., Jaton, K. , and Opota, O. (2020). Computer-aided medical microbiology monitoring tool: a strategy to adapt to the SARS-CoV-2 epidemic and that highlights RT-PCR consistency. *medRxiv* 2020.2007.2027.20162123. doi: 10.1101/2020.07.27.20162123

Mutters, N. T., Hodiamont, C. J., de Jong, M. D., Overmeijer, H. P. J., van den Boogaard, M., and Visser, C. E. (2014). Performance of Kiestra total laboratory automation combined with MS in clinical microbiology practice. *Ann. Lab. Med.* 34 (2), 111–117. doi: 10.3343/alm.2014.34.2.111

Muzzi, A., Brozzi, A., Serino, L., Bodini, M., Abad, R., Caugant, D., et al. (2019). Genetic Meningococcal Antigen Typing System (gMATS): A genotyping tool that predicts 4CMenB strain coverage worldwide. *Vaccine* 37 (7), 991–1000. doi: 10.1016/j.vaccine.2018.12.061

Olearo, F., Marinosci, A., Stephan, R., Cherkaoui, A., Renzi, G., Gaia, N., et al. (2020). First case of Streptococcus suis infection in Switzerland: An emerging public health problem? *Travel Med. Infect. Dis.* 36, 101590. doi: 10.1016/j.tmaid.2020.101590

Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 16 (1), 236. doi: 10.1186/s12864-015-1419-2

Patel, R. (2019). A Moldy Application of MALDI: MALDI-ToF Mass Spectrometry for Fungal Identification. *J. Fungi (Basel)* 5 (1), 4. doi: 10.3390/jof5010004

Peiffer-Smadja, N., Dellière, S., Rodriguez, C., Birgand, G., Lescure, F. X., Fourati, S., et al. (2020). Machine learning in the clinical microbiology laboratory: has the time come for routine practice? *Clin. Microbiol. Infect.* 26 (10), 1300–1309 doi: 10.1016/j.cmi.2020.02.006

Pérez-Losada, M., Arenas, M., and Castro-Nallar, E. (2018). Microbial sequence typing in the genomic era. *Infect. Genet. Evol.* 63, 346–359. doi: 10.1016/j.meegid.2017.09.022

Pham, T.-T., Lazarevic, V., Gaia, N., Girard, M., Cherkaoui, A., Suva, D., et al. (2020). Second periprosthetic joint infection caused by Streptococcus dysgalactiae: how genomic sequencing can help defining the best therapeutic strategy. *Front. Med.* 7 (53). doi: 10.3389/fmed.2020.00053

Ruppé, E., and Schrenzel, J. (2018). Messages from the second International Conference on Clinical Metagenomics (ICCMg2). *Microbes Infect.* 20 (4), 222–227. doi: 10.1016/j.micinf.2018.02.005

Ruppé, E., and Schrenzel, J. (2019). Messages from the third International Conference on Clinical Metagenomics (ICCMg3). *Microbes Infect.* 21 (7), 273–277. doi: 10.1016/j.micinf.2019.02.004

Ruppé, E., Baud, D., Schicklin, S., Guigon, G., and Schrenzel, J. (2016). Clinical metagenomics for the management of hospital- and healthcare-acquired pneumonia. *Future Microbiol.* 11 (3), 427–439. doi: 10.2217/fmb.15.144

Ruppé, E., Greub, G., and Schrenzel, J. (2017). Messages from the first International Conference on Clinical Metagenomics (ICCMg). *Microbes Infect.* 19 (4–5), 223–228 doi: 10.1016/j.micinf.2017.01.005

Scherrer, S., Rosato, G., Spoerry Serrano, N., Stevens, M. J. A., Rademacher, F., Schrenzel, J., et al. (2020). Population structure, genetic diversity and pathotypes of Streptococcus suis isolated during the last 13 years from diseased pigs in Switzerland. *Vet. Res.* 51 (1), 85. doi: 10.1186/s13567-020-00813-w

Shin, J. H. (2018). "Nucleic Acid Extraction and Enrichment," in *Advanced Techniques in Diagnostic Microbiology: Volume 1: Techniques*. Eds. Y.-W. Tang and C. W. Stratton (Cham: Springer International Publishing), 273–292. doi: 10.1007/978-3-319-33900-9_13

Smith, K. P., Richmond, D. L., Brennan-Krohn, T., Elliott, H. L., and Kirby, J. E. (2017). Development of MAST: a microscopy-based antimicrobial susceptibility testing platform. *SLAS Technol.* 22 (6), 662–674. doi: 10.1177/2472630317727721

Smith, K. P., Kang, A. D., and Kirby, J. E. (2018). Automated interpretation of blood culture Gram stains by use of a deep convolutional neural network. *J. Clin. Microbiol.* 56 (3), e01521–e01517. doi: 10.1128/JCM.01521-17

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25 (1), 44–56. doi: 10.1038/s41591-018-0300-7

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903. doi: 10.1038/nmeth.3589

Vandenberg, O., Kozlakidis, Z., Schrenzel, J., Struelens, M. J., and Breuer, J. (2018). Control of infectious diseases in the era of European clinical microbiology laboratory consolidation: new challenges and opportunities for the patient and for public health surveillance. *Front. Med.* 19 (4-5), 223–228. doi: 10.3389/fmed.2018.00015

Vandenberg, O., Durand, G., Hallin, M., Diefenbach, A., Gant, V., Murray, P., et al. (2020). Consolidation of clinical microbiology laboratories andintroduction of transformative technologies. *Clin. Microbiol. Rev.* 33 (2), e00057–e00019. doi: 10.1128/CMR.00057-19

Wagner, J., Paulson, J. N., Wang, X., Bhattacharjee, B., and Corrada Bravo, H. (2016). Privacy-preserving microbiome analysis using secure computation. *Bioinformatics* 32 (12), 1873–1879. doi: 10.1093/bioinformatics/btw073

Wan, X.-F., Barnett, J. L., Cunningham, F., Chen, S., Yang, G., Nash, S., et al. (2013). Detection of African swine fever virus-like sequences in ponds in the Mississippi Delta through metagenomic sequencing. *Virus Genes* 46 (3), 441–446. doi: 10.1007/s11262-013-0878-2

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3 (1), 160018. doi: 10.1038/sdata.2016.18

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15 (3), R46. doi: 10.1186/gb-2014-15-3-r46

Xu, B., Liu, L., Huang, X., Ma, H., Zhang, Y., Du, Y., et al. (2011). Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. *PloS Pathog.* 7 (11), e1002369. doi: 10.1371/journal.ppat.1002369

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67 (11), 2640–2644. doi: 10.1093/jac/dks261

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2020 Leo, Cherkaoui, Renzi and Schrenzel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 08 January 2021
doi: 10.3389/fcimb.2020.594030

# Machine Learning Algorithms Evaluate Immune Response to Novel *Mycobacterium tuberculosis* Antigens for Diagnosis of Tuberculosis

*Noëmi Rebecca Meier[1,2], Thomas M. Sutter[3], Marc Jacobsen[4], Tom H. M. Ottenhoff[5], Julia E. Vogt[3] and Nicole Ritz[1,2,6,7*†] on behalf of the CITRUS Study Team[†]*

[1] Mycobacterial Research Laboratory, University of Basel Children's Hospital, Basel, Switzerland, [2] Faculty of Medicine, University of Basel, Basel, Switzerland, [3] Department of Computer Science, Medical Data Science, Eidgenössische Technische Hochschule (ETH) Zurich, Zurich, Switzerland, [4] Department of General Pediatrics, Neonatology and Pediatric Cardiology, University Children's Hospital, Heinreich Heine University, Düsseldorf, Germany, [5] Department of Infectious Diseases, Leiden University Medical Center, Leiden, Netherlands, [6] Pediatric Infectious Diseases and Vaccinology Unit, University of Basel Children's Hospital, Basel, Switzerland, [7] Department of Pediatrics, Royal Children's Hospital Melbourne, University of Melbourne, Parkville, VIC, Australia

OPEN ACCESS

**Edited by:**
Adrian Egli,
University Hospital of Basel,
Switzerland

**Reviewed by:**
Hirdesh Kumar,
National Institutes of Health (NIH),
United States
Charles William Stratton,
Vanderbilt University Medical Center,
United States

***Correspondence:**
Nicole Ritz
nicole.ritz@unibas.ch

[†]Corporate authorship

**Specialty section:**
This article was submitted to Clinical Microbiology, a section of the journal Frontiers in Cellular and Infection Microbiology

**Received:** 12 August 2020
**Accepted:** 24 November 2020
**Published:** 08 January 2021

**Citation:**
Meier NR, Sutter TM, Jacobsen M, Ottenhoff THM, Vogt JE and Ritz N (2021) Machine Learning Algorithms Evaluate Immune Response to Novel Mycobacterium tuberculosis Antigens for Diagnosis of Tuberculosis. Front. Cell. Infect. Microbiol. 10:594030. doi: 10.3389/fcimb.2020.594030

**Rationale:** Tuberculosis diagnosis in children remains challenging. Microbiological confirmation of tuberculosis disease is often lacking, and standard immunodiagnostic including the tuberculin skin test and interferon-$\gamma$ release assay for tuberculosis infection has limited sensitivity. Recent research suggests that inclusion of novel *Mycobacterium tuberculosis* antigens has the potential to improve standard immunodiagnostic tests for tuberculosis.

**Objective:** To identify optimal antigen–cytokine combinations using novel *Mycobacterium tuberculosis* antigens and cytokine read-outs by machine learning algorithms to improve immunodiagnostic assays for tuberculosis.

**Methods:** A total of 80 children undergoing investigation of tuberculosis were included (15 confirmed tuberculosis disease, five unconfirmed tuberculosis disease, 28 tuberculosis infection and 32 unlikely tuberculosis). Whole blood was stimulated with 10 novel *Mycobacterium tuberculosis* antigens and a fusion protein of early secretory antigenic target (ESAT)-6 and culture filtrate protein (CFP) 10. Cytokines were measured using xMAP multiplex assays. Machine learning algorithms defined a discriminative classifier with performance measured using area under the receiver operating characteristics.

**Measurements and main results:** We found the following four antigen–cytokine pairs had a higher weight in the discriminative classifier compared to the standard ESAT-6/CFP-10-induced interferon-$\gamma$: Rv2346/47c- and Rv3614/15c-induced interferon-gamma inducible protein-10; Rv2031c-induced granulocyte-macrophage colony-stimulating factor and ESAT-6/CFP-10-induced tumor necrosis factor-$\alpha$. A combination of the 10 best antigen–cytokine pairs resulted in area under the curve of 0.92 ± 0.04.

**Conclusion:** We exploited the use of machine learning algorithms as a key tool to evaluate large immunological datasets. This identified several antigen–cytokine pairs with the potential to improve immunodiagnostic tests for tuberculosis in children.

**Keywords: cytokines, novel antigens, immune response, pediatric tuberculosis, interferon-gamma release assay**

## INTRODUCTION

Tuberculosis (TB) remains one of the leading causes of death globally. Current estimates show that one in ten TB cases occur in children below 15 years of age with an annual estimated number of one million cases of childhood TB disease in 2017 (World Health Organization, 2018a). Despite being a preventable and curable disease, 233,000 children died of TB in 2017, of which 80% occurred in children below 5 years of age. The recent World Health Organization roadmap towards ending TB in children and adolescents mentions up to 69% underdiagnosis and highlights the development of accurate, non-sputum-based diagnostics tests for TB disease and infection as a key action towards ending TB in children and adolescents (World Health Organization, 2018b).

TB infection is characterized by the absence of clinical signs and symptoms and evidence of containment of disease through the host immunological response. TB disease is usually defined as the active state of disease with loss of immunological containment, presence of symptoms and risk of transmission of disease. In young children TB disease is often of paucibacillary nature (*i.e.* low mycobacterial bacterial load) and therefore may remain undiagnosed using microbiological assays (Perez-Velez and Marais, 2012). In addition collection of samples for microbiological proof in this patient group is challenging and therefore TB confirmation reaches 50% at best (Oesch Nemeth et al., 2014). As a consequence, non-sputum-based diagnostic tests based on immunological evidence of TB have been developed. These tests rely on the measurement of a recall cell mediated immune response triggered by *in vivo* or *in vitro* mycobacterial antigens. Until two decades ago the tuberculin skin test has been the standard test, measuring a local skin induration after injection of purified protein derivative, a *Mycobacterium tuberculosis* protein mixture. However, because of its low specificity especially in Bacille Calmette–Guérin (TB vaccine prepared from an attenuated strain of *Mycobacterium bovis*) vaccinated individuals, interferon-gamma release assays have been developed, and have become the standard immunodiagnostic test of TB infection in adults (Diel et al., 2010). Interferon-gamma release assays are *in-vitro* blood-based assays measuring the *Mycobacterium tuberculosis*-specific immune response. Unfortunately these assays have two major limitations: lower performance in children with a sensitivity ranging from 62 to 83% and inability to discriminate between TB disease and TB infection (Mandalakas et al., 2011; Sollai et al., 2014). Recent research suggests that incorporation of novel *Mycobacterium tuberculosis* antigens expressed during different stages of TB [reviewed in (Meier et al., 2018)] and the measurement of additional cytokines (Walzl et al., 2011) can improve performance of currently used interferon-gamma release assay. Evaluation of novel diagnostic tests incorporating different *Mycobacterium tuberculosis* antigens and cytokines is therefore a feasible test suitable for pediatrics and urgently needed (World Health Organization, 2013).

The aim of our study was to include novel *Mycobacterium tuberculosis* antigens and measure additional cytokines for the immune diagnosis of childhood TB. We used supervised and unsupervised machine learning algorithms to compare groups and identify the best antigen–cytokine pairs.

**Abbreviations:** AUROC, area under the receiver operating characteristic; CFP-10, 10 kDA culture filtrate protein; CITRUS, Childhood Tuberculosis in Switzerland Study; ESAT-6, early secretory antigenic target 6; GM-CSF, granulocyte-macrophage colony-stimulating factor; IFN, interferon; IL, interleukin; IP-10, interferon-gamma inducible protein 10; min–max, minimum–maximum; mean–std, mean – standard deviation; sCD40L, soluble cluster of differentiation 40 ligand; TB, tuberculosis; TNF, tumor necrosis factor.

## METHODS

### Study Design, Setting, and Population

The Childhood Tuberculosis in Switzerland Study (CITRUS) is a prospective multicenter observational study (registered at ClinicalTrials.gov NCT03044509 and approved by the ethics committee EKNZ 2016-01094). In brief, eligible are children undergoing evaluation for TB exposure, infection or disease below the age of 18 years. Children that have been treated previously or that have started treatment more than 5 days before study inclusion are excluded. Upon enrolment baseline characteristics, clinical scores and TB test results done by the treating physician are recorded. The study participants were classified into the following groups confirmed TB, unconfirmed TB, TB infection, unlikely TB according to previously published case definitions (Graham et al., 2015) (for further details on study design and population see **Supplementary Methods Text**).

### Sample Preparation and Stimulation

Blood was collected in lithium-heparin tubes (Sarstedt Monovette 01.1608.100) and stimulated within 8 h of collection with 5 μg/ml phytohaemagglutinin (Merck chemicals LTD., Beeston, Nottingham, UK), 10 μg/ml staphylococcus enterotoxin B (Sigma Aldrich GmbH, Schnelldorf, Germany), 5 μg/ml of the following *Mycobacterium tuberculosis* recombinant proteins expressed and purified in *Escherichia coli* BL21: Rv0081, Rv1733c, Rv2031c, Rv0867c, Rv2389c, Rv3407, Rv2346/47c, Rv2431c, Rv3614/15c, Rv3865 and a fusion protein of early secretory antigenic target 6 (ESAT-6) and 10 kDa culture filtrate protein (CFP-10) [provided by the Department of

Infectious Diseases at the University Leiden, the Netherlands (Franken et al., 2000),] and an unstimulated control (no protein added). The selection of the *Mycobacterium tuberculosis* recombinant proteins was based on published data summarized in a systematic literature review (Meier et al., 2018) and from unpublished data (personal communication THM Ottenhoff) CD28 and CD49d antibodies (Biolegend Inc., San Diego, Ca 92121, USA) were added at a concentration of 1 µg/ml to all conditions. Samples were stimulated overnight (16–18 h) at 37°C (**Figure 1A**).

## Cytokine Measurement

Granulocyte-macrophage colony-stimulating factor (GM-CSF), interferon (IFN)-γ, IFN-γ-inducible protein (IP)-10, interleukin (IL)-1 receptor-antagonist (RA), IL-2, IL-6, IL-10, IL-13, IL-17, soluble cluster of differentiation 40 ligand (sCD40L) and tumor necrosis factor (TNF)-α were measured using a Luminex technology according to the manufacturer's instructions (**Figure 1A**, **Supplementary Methods Text**).

## Normalization of Data

Cytokine concentrations were normalized (Dodge, 2006) within antigen–cytokine pairs (using a minimum–maximum (min–max) or a mean-standard deviation (mean–std) normalization) and within a patient's distribution of values (using a mean–std normalization) as indicated (**Figure 1B**).

## Discriminative Classifier

Discrimination of a pre-defined binary outcome (confirmed/unconfirmed TB and TB infection *versus* TB exposed), based on data containing information on all antigen–cytokine pairs (features), was achieved using a logistic regression classifier with L2-regularization (Hoerl and Kennard, 1970) (**Supplementary Methods Text**). To get a reliable estimate of the discriminative classifier performance, a five-fold cross-validation was applied to a set of training data to select the model's hyperparameters (see **Supplementary Methods**). The performance of the discriminative classifier was evaluated using area under the receiver operating characteristics (AUROC)

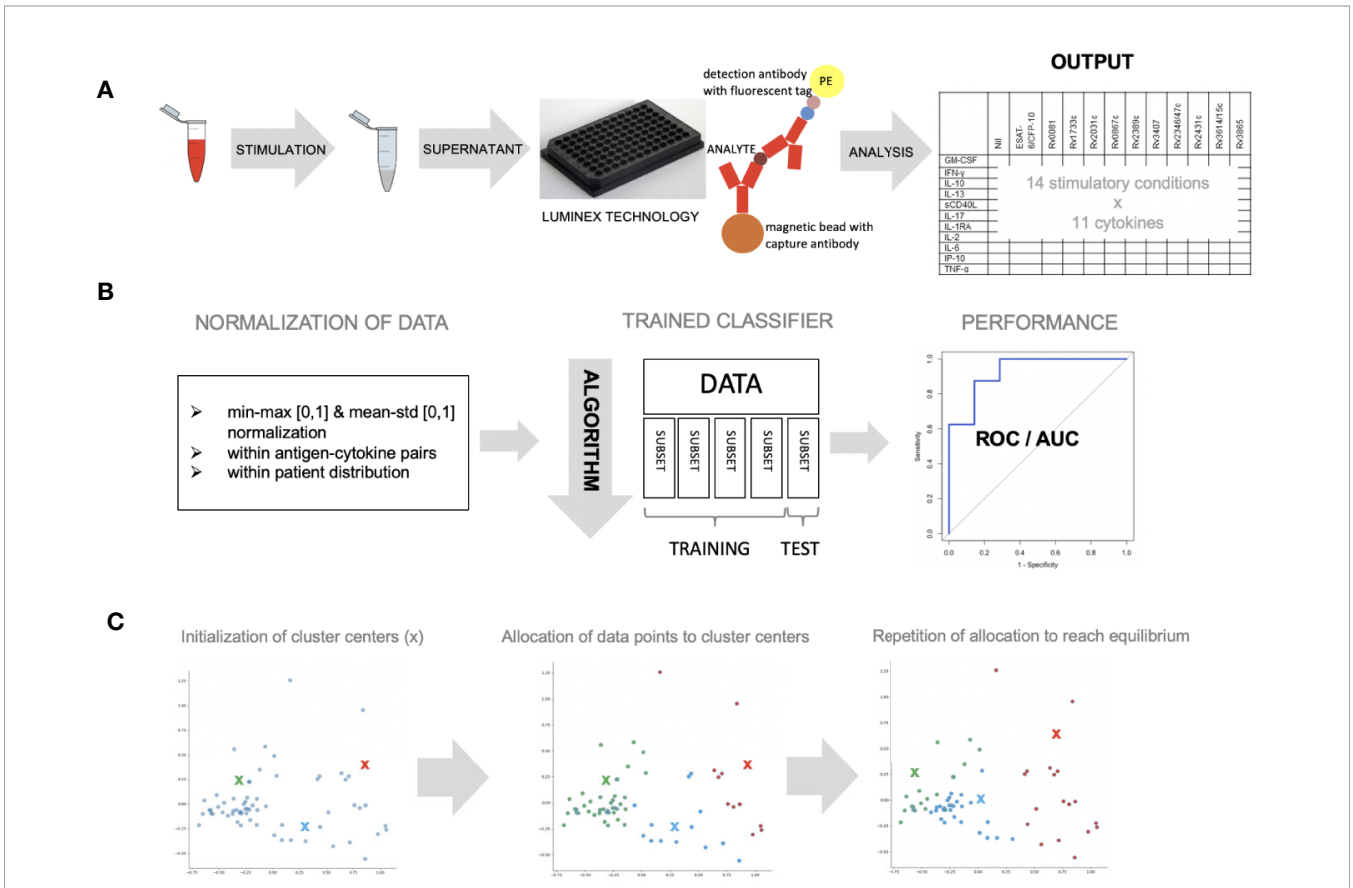



**FIGURE 1** | Whole blood was stimulated with novel antigens and data was analyzed with different machine learning algorithms. **(A)** Whole blood was stimulated with 11 mycobacterial antigens, left unstimulated and with a positive control, overnight and supernatant was analyzed using Luminex technology to measure 11 different cytokines **(B)** data was normalized within antigen–cytokine pairs using min-max or mean–std normalization or within patient distribution using the latter only. Data (n = 59) was divided into five equal parts and a classifier discriminating healthy *vs.* sick children was trained using four subsets and tested on one subset (cross-validation). The algorithm's parameters were adjusted until performance was optimal. ROC curves were used to measure performance. **(C)** K-means clustering approach was used to allocate individual data points to three cluster centers randomly. This approach was repeated until optimal data point allocation was reached meaning the sum of the distances from data point to cluster centers is minimized.

(Hanley and Mcneil, 1982). The contribution of each antigen–cytokine pair to our predictive model was evaluated by analyzing the weight in the decision function (**Figure 1B**).

## Unsupervised K-Means Clustering

K-means clustering algorithm (MacQueen, 1967) was performed with a predefined number of clusters (n = 3) reflecting the anticipated number of patient groups (confirmed/unconfirmed TB disease, TB infection, unlikely TB). Patients with incomplete measurements in any of the conditions (*e.g.* missing values) were excluded from this analysis. Cluster centers were allocated randomly at first, and every patient was then assigned to the nearest cluster center. Cluster center allocation and data point assignment were repeated until an equilibrium was reached (sum of distances is minimized, cluster centers not changed) (**Figure 1C**).

## Supervised K-Means Clustering Based on Median Cytokine Differences

Differences in median cytokine concentrations between confirmed/unconfirmed TB, TB infection and unlikely TB were compared. Antigen–cytokine pairs with the greatest differences were selected and K-means clustering approach was performed as above on these selected antigen–cytokine pairs.

# RESULTS

A total of 80 patients were included: confirmed TB disease (n = 15), unconfirmed TB disease (n = 5), TB infection (n = 28), and unlikely TB (n = 32). Median age in the three TB groups was as follows: 9.7, 12.0, 11.3, and 5.8 years for confirmed TB, unconfirmed TB, TB infection, and unlikely TB (**Table 1**). A total of 49 of 80 (61.3%) children were tested for HIV, and all were negative. A total of 39 study participants out of 80 were born in Switzerland (48.8%), and 31 of 80 (38.8%) arrived in Switzerland less than 3 years prior to inclusion to the study. Routine immunodiagnostic testing was performed in 77 children with QuantiFERON-TB in 57/77 (74.0%) children, T-SPOT.TB in 10/77 (13.0%) and a tuberculin skin test in 40/77 (51.9%) children. Both interferon-gamma release assay and tuberculin skin test were done in 30 children and showed 23 (76.7%) concordant and 7 (23.3%) discordant results (one QuantiFERON-TB +/tuberculin skin test-; six QuantiFERON-TB −/tuberculin skin test+). Two T-SPOT.TB results were indeterminate (a confirmed TB disease case and an unconfirmed TB disease case).

## A Discriminative Classifier Distinguishes Healthy From Sick Children and Normalization of Data Results in Improvement of the Classifier's Performance

A total of 59 patients had complete measurements for all antigen–cytokine pairs and were included in this analysis: confirmed TB (n = 8), unconfirmed TB (n = 2), TB infection (n = 17) and unlikely TB (n = 32). Different methods of normalization (*e.g.* non-normalized data, antigen–cytokine pairs either normalized using min–max or mean–std normalization and normalization of antigen–cytokine pairs with min–max and between patient normalization with mean–std were applied to our dataset and resulted in differences on visual inspection of the graphs between antigen–cytokine pairs and cytokine concentrations (**Supplementary Figures S1A–D**). These differences influenced the outcome of the discriminative classifier (confirmed/unconfirmed TB and TB infection *versus* TB exposed). The AUROC was lower without normalization (AUROC = 0.81 ± 015), compared to a normalization of antigen–cytokine pairs ($AUROC_{min-max}$ = 0.89 ± 0.12 and $AUROC_{mean-std}$ = 0.87 ± 0.13) or combining an antigen–cytokine pair normalization with individual patient normalization ($AUROC_{min-max/mean-std}$ = 0.95 ± 0.03) (**Figure 2B**). The most important antigen–cytokine pairs that contributed to the performance of the discriminative classifier were consistent for the normalization methods used. Rv2346/47c- and Rv3614/15c-induced concentrations of IP-10 were the two antigen–cytokine pairs with the highest weight in the predictive model for all discriminative classifiers with normalized data (**Figure 3B**, **Supplementary Figures S2A–C**). The weight of ESAT-6 and CFP-10-induced concentrations of TNF-$\alpha$ for the predictive model was consistently high for all normalized and non-normalized data. ESAT-6/CFP-10-induced concentrations of IFN-$\gamma$ were among the 10 antigen–cytokine pairs that contributed the most to the classifier for all non-normalized and normalized data except when mean–std normalization alone was applied. Rv2031c-induced concentrations of GM-CSF contributed to the performance of the classifier when any normalization method was applied with increasing weight for combined min–max and mean–std normalization. Combining data from the 10 antigen–cytokine pairs with the highest weight in the predictive model using both min–max and mean–std normalization resulted in $AUROC_{min-max/mean-std}$ = 0.92 ± 0.04 (**Figure 3A**).

## Unsupervised K-Means Clustering Reveals Three Groups of Children That Cannot Be Explained by Disease Status

K-means is a machine learning tool using vector quantization that groups observations into clusters based on distances to allocated cluster centers. Thereby we found three clusters which did not overlap with our patient groups (*i.e.* confirmed and unconfirmed TB, TB infection, unlikely TB) in the unsupervised analysis approach. All three clusters included patients from all study groups. **Figure 2A** displays normalized cytokine concentrations of antigen–cytokine pairs of all individual patients sorted by cluster (2, 1 or 0). Cluster 0 consisted of four confirmed TB, one unconfirmed TB, six TB infection and five unlikely TB patients (median age = 8.4, 68.7% male). Cluster 1 consisted of two confirmed TB, zero unconfirmed TB, two TB infection, and one unlikely TB patients (median age = 13.6, 20.0% male). Cluster 2 consisted of two confirmed TB, one unconfirmed TB, nine TB infection and 26 unlikely TB patients (median age = 7.8, 55.3% male). Clusters could neither be explained by disease classification, nor age, nor gender, nor ethnicity (data not shown).

**TABLE 1 |** Baseline characteristics of the study population according to study group.

| Variable | Confirmed TB | Unconfirmed TB | TB infection | Unlikely TB | Total |
|---|---|---|---|---|---|
| | N = 15 | N = 5 | N = 28 | N = 32 | N = 80 |
| Median age, range (years) | 9.7 (0.9–15.9) | 12 (3–15.8) | 11.3 (0.2–17.1) | 5.8 (0.2–16.7) | 9.6 (0.2–17.1) |
| IQR age | 3.1–15.2 (12.1) | 9.6–15.4 (5.8) | 8.1–13.5 (5.4) | 3.0–10.2 (7.2) | 3.5–12.8 (9.3) |
| Males | 6 (40%) | 2 (40%) | 15 (53.6%) | 19 (59.4%) | 42 (52.5%) |
| Median weight, range (m) | 38.7 (11–75) | 49 (13–60) | 44 (10–71) | 20 (8–65) | 33.5 (8–75) |
| Ethnicity | | | | | |
| Caucasian | 2 | 2 | 11 | 10 | 25 |
| African | 9 | 2 | 9 | 11 | 31 |
| Asian | – | – | 2 | 6 | 8 |
| other | 4 | 1 | 6 | 5 | 16 |
| Country of birth | | | | | |
| Born in Switzerland | 6 | 3 | 10 | 20 | 39 |
| Recently migrated to Switzerland | 9 | 2 | 11 | 9 | 31 |
| Unknown | 1 | 0 | 5 | 4 | 10 |
| Symptoms | | | | | |
| Asymptomatic | 5 | 1 | 25 | 29 | 60 |
| Symptoms | 10 | 4 | 3 | 2 | 19 |
| cough | 8 | 3 | 1 | 2 | 14 |
| fever | 6 | 2 | 0 | 2 | 10 |
| unexplained fatigue | 4 | 0 | 0 | 0 | 4 |
| weight loss | 4 | 2 | 0 | 0 | 6 |
| lack of weight gain | 1 | 0 | 0 | 0 | 1 |
| other symptoms | 3 | 4 | 1 | 1 | 9 |
| Bacille Calmette-Guérin vaccination status | | | | | |
| vaccinated | 4 | 0 | 9 | 8 | 21 |
| not vaccinated | 5 | 3 | 9 | 21 | 38 |
| unknown | 6 | 2 | 10 | 3 | 21 |
| HIV status | | | | | |
| negative | 12 | 5 | 14 | 18 | 49 |
| positive | 0 | 0 | 0 | 0 | 0 |
| unknown | 3 | 0 | 14 | 14 | 31 |
| Tuberculin skin test | | | | | |
| not done | 10 | 5 | 15 | 10 | 40 |
| <5 mm | 0 | 0 | 1 | 19 | 20 |
| >5 mm | 2 | 0 | 7 | 3 | 12 |
| >15 mm | 3 | 0 | 5 | 0 | 8 |
| Imaging | | | | | |
| X ray | 15 | 5 | 27 | 24 | 71 |
| CT | 8 | 4 | 3 | 1 | 16 |
| compression | 2 | 1 | 0 | 0 | 3 |
| lymphadenopathy | 9 | 2 | 0 | 0 | 11 |
| consolidation parenchyma | 11 | 4 | 0 | 2 | 17 |
| miliary pattern | 0 | 0 | 0 | 0 | 0 |
| pleural effusion | 3 | 0 | 0 | 0 | 3 |
| cavitation | 4 | 3 | 0 | 0 | 7 |
| TSPOT | | | | | |
| not done | 13 | 4 | 24 | 29 | 70 |
| negative | 0 | 0 | 0 | 3 | 3 |
| positive | 2 | 1 | 4 | 0 | 7 |
| QuantiFERON-TB | | | | | |
| not done | 5 | 2 | 5 | 11 | 23 |
| negative | 1 | 1 | 4 | 21 | 27 |
| positive | 9 | 2 | 19 | 0 | 30 |
| Microbiological confirmation | | | | | |
| not done | 0 | 0 | 22 | 30 | 52 |
| culture positive | 13 | 0 | 0 | 0 | 13 |
| culture negative | 0 | 4 | 6 | 2 | 12 |
| PCR positive | 13 | 0 | 0 | 0 | 13 |
| PCR negative | 0 | 5 | 4 | 0 | 9 |

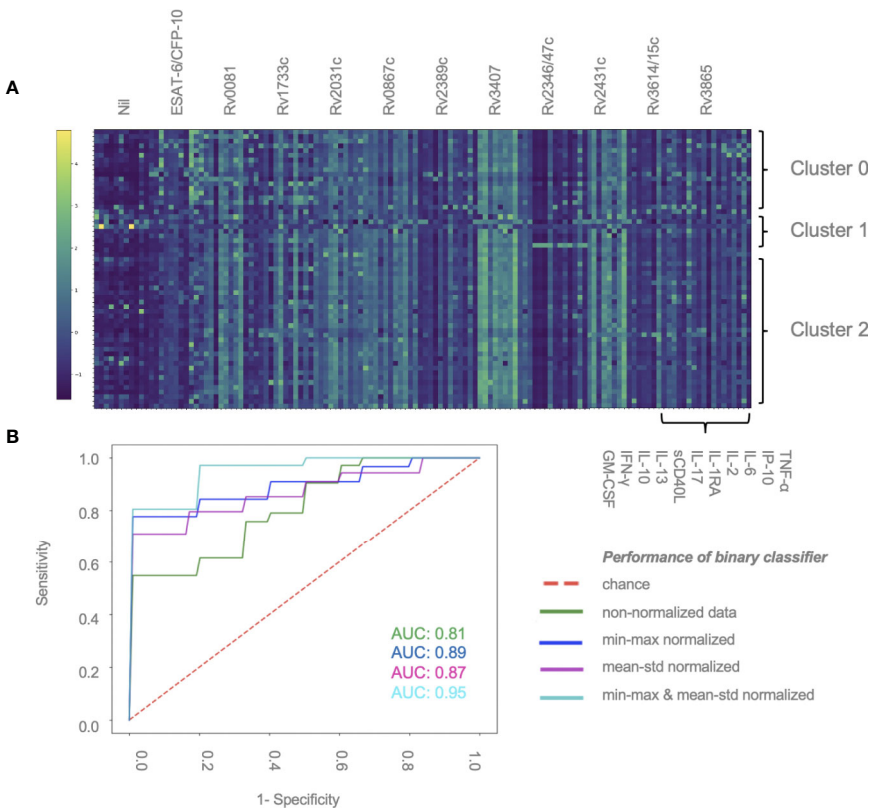



**FIGURE 2** | Normalization of data contributes to performance of discriminative classifier **(A)** Cytokine concentrations for individual patients. Results are sorted by patient group and clusters (2, 1 or 0), and antigen–cytokine pairs. Clustering was performed using K-means algorithm. Min–max normalization was applied to cytokine–antigen concentrations, mean–std normalization was applied to between-individual measurements (color change from dark blue to light green represents an increase in relative cytokine concentration). **(B)** AUROC curve showing the performance of the binary classifier (confirmed/unconfirmed TB and TB infection *versus* TB exposed) in 59 patients using different normalization methods: min–max and mean–std; normalization of antigen–cytokine pairs; min–max/mean-std combining an antigen–cytokine pair normalization with individual patient normalization.

## Supervised K-Means Clustering Based on Median Cytokine Differences Between Three Study Groups Reveals One Group That Clustered Mainly Healthy Children but No Confirmed TB Cases

Greatest differences in median cytokine concentrations between confirmed/unconfirmed TB, TB infection and unlikely TB were observed for: ESAT-6/CFP-10-induced concentrations of GM-CSF, IFN-$\gamma$ and IL-2; Rv0081-induced concentrations of TNF-$\alpha$; Rv2389c-induced concentrations of GM-CSF and IP-10; and Rv3614/15c-induced concentrations of IFN-$\gamma$, IL-2, IP-10 and TNF-$\alpha$ (data not shown). A total of 71 patients had complete measurements for these 10 conditions with the greatest differences and were thus further included in the comparative analysis: confirmed TB (n = 10), unconfirmed TB (n = 4), TB infection (n = 25) and unlikely TB (n = 32). K-means clustering with these antigen–cytokine pairs resulted in three cluster grouping the majority of unlikely TB patients and none of the confirmed TB patients in cluster 0 (25 out of 32). Only one unlikely TB patient and none of the unconfirmed TB patients were grouped to cluster 2 (six confirmed TB, five TB infection). Cluster 1 consisted of all four study groups with the majority being TB infected (11 out of 24) (**Supplementary Figures S3A–B**).

## DISCUSSION

Diagnosis of childhood TB is one of the key challenges for the global epidemic. As current diagnostic tests are insufficient for detection of TB in children, there is an urgent need for novel tests. Our study is unique as it combines the use of the largest number of novel *Mycobacterium tuberculosis* antigens and cytokine combinations in a childhood TB diagnostic study, exploring the results by applying different machine learning algorithms.

We found that IP-10-responses induced by Rv2346/47c and Rv3614/15c were the two most important features to discriminate diseased from healthy individuals. We showed that further cytokines including GM-CSF, IL-2, IL-6, INF-$\gamma$ and TNF-$\alpha$ play an important role during immune responses in TB in children. We also demonstrate the importance of data normalization to reduce bias towards highly expressed cytokines and inter-individual heterogeneity in *Mycobacterium tuberculosis*-specific immune responses.

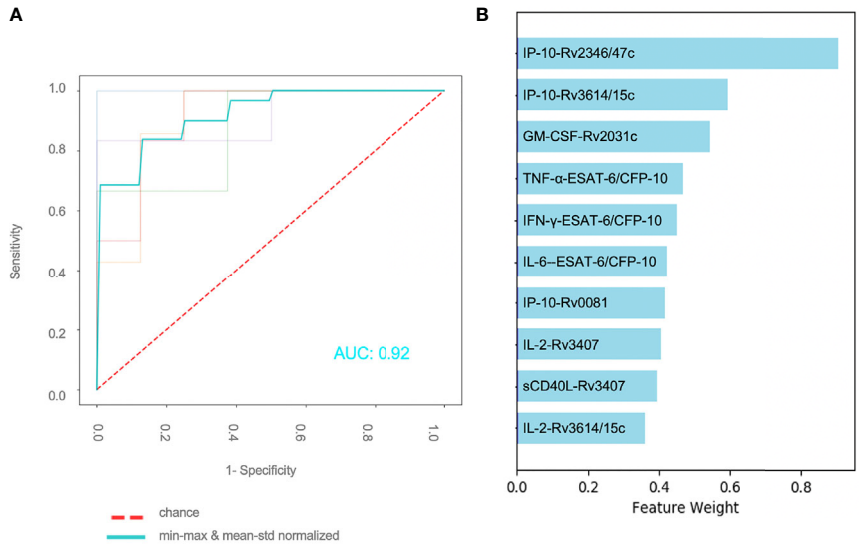



**FIGURE 3** | Effect of normalization of antigen–cytokine pairs and normalization for individual patients **(A)** Performance of binary classifier using the 10 most important features and applying an antigen–cytokine pair normalization (min–max) and a normalization for individual patients (mean–std) **(B)** Combination of 10 antigen–cytokine pairs contributing the most to performance of trained discriminative classifier with min–max normalization of antigen–cytokine pairs and mean–std individual patient normalization.

Our selection of novel *Mycobacterium tuberculosis* antigens was based on previously published studies, and the antigens that are expressed during different stages of TB are briefly summarized below. The dormancy of survival regulon encoded antigens (Rv0081, Rv1733c, and Rv2031c) belong to a region of the *Mycobacterium tuberculosis* genome that includes approximately 50 genes associated with the non-replicative stage of TB (Voskuil et al., 2003). These antigens together with reactivation associated antigens (Rv0867c, Rv2389c, Rv3407) are highly immunogenic and have been tested mainly in adult cohorts [reviewed in (Meier et al., 2018)]. We also included the recently discovered *in vivo*-expressed antigens (Rv2346/47c, Rv2431c, Rv3614/15c, Rv3865) that have not been studied extensively in humans, but are believed to be important virulence factors (Commandeur et al., 2013). Rv2346 and Rv2347c are ESAT-6 like proteins and associated with downregulation of IL-6 and TNF-α enabling survival of bacteria inside macrophages (Malen et al., 2007; Yao et al., 2018). Rv2431c is a prolin-glutamic acid family protein, and its function is yet to be understood (Malen et al., 2007). Previous studies showed its involvement in necrosis in macrophages (Tundup et al., 2014) but also maturation and proliferation of dendritic cells (Chen et al., 2016). The antigens Rv3614c, Rv3615c and Rv3865 are all associated with the ESAT-6 secretion system 1 absent in the Bacille Calmette–Guérin vaccine strains.

The diagnostic potential of the recently discovered *in vivo*-expressed antigens found in our study has been shown in previous studies confirming our results (Millington et al., 2011). IFN-$\gamma$ responses induced by Rv3615c were as specific as ESAT-6 and CFP-10 induced IFN-$\gamma$ responses in patients with TB disease and infection (Millington et al., 2011). The antigen Rv3615c was included in a modified T-SPOT.TB assay and was shown to improve the diagnosis of TB disease and infection compared to healthy controls and patients with non-TB lung disease (Li et al., 2017). The use of Rv3865 seems to be of limited value also shown by the low immunogenic potential in other studies in adults (Bahk et al., 2004) and adolescents including different stages of TB infection (Michelsen et al., 2017).

In our study we found dormancy of survival regulon encoded antigens to be of key importance eliciting a differential immune response in TB patients and exposed healthy controls. We found that the dormancy of survival regulon antigens Rv0081 and Rv2031c-induced IP-10 and GM-CSF responses contributed strongly to performance of the discriminative classifier. Several studies in adults reported elevated concentrations of cytokines induced by Rv0081 during TB infection and disease, which is in line with our findings [reviewed in (Meier et al., 2018)]. In contrast to our findings, studies in adults suggest Rv1733c-induced immune responses to be of added diagnostic value (Leyten et al., 2006; Kassa et al., 2012; Mensah et al., 2014; Serra-Vidal et al., 2014). Furthermore, previous studies including Rv2031c-induced cytokine response, showed conflicting results with one study reporting higher concentrations of IFN-$\gamma$, IL-10, and TNF-α in TB exposed individuals compared to healthy controls (Belay et al., 2015) and other studies failing to show IFN-$\gamma$ responses induced by this antigen (Goletti et al., 2010; Hozumi et al., 2013). Our study supports the notion that Rv2031c-induced responses are important as diagnostic markers for TB particularly when cytokines other than IFN-$\gamma$ are included into the analysis. This is in line with Coppola et al. showing high concentrations of TNF-α expression in response to Rv2031c in addition to other cytokines such as IP-10 or IL-17 but notably not IFN-$\gamma$ (Coppola et al., 2016).

In addition to the above, two reactivation-associated antigens were found to be important in our study: Rv3407 and Rv2389c. We found that Rv2389c-induced GM-CSF and IP-10 responses were among the 10 antigen–cytokine pairs that contributed the most to discriminating between sick and healthy. Other studies also show the diagnostic potential of Rv2389c. IFN-$\gamma$ responses induced by Rv0867c and Rv2389c were found to be higher in individuals with TB infection compared to healthy controls and TB disease in several studies (Commandeur et al., 2011; Chegou et al., 2012; Serra-Vidal et al., 2014). High concentrations of IL-6, IL-10, and TNF-$\alpha$ were found to be induced by Rv0867c and Rv2389c in individuals with TB disease (Kassa et al., 2012). In our study, however, Rv0867c did not induce cytokine responses that contributed to classification of patients.

The standard antigens used in the current available test including ESAT-6 and CFP-10 remain important. Our results, however, clearly show that in addition to IFN-$\gamma$ also IL-6 and TNF-$\alpha$ responses to ESAT-6 and CFP-10 contributed towards distinction of study groups and were among the 10 most important features for the discriminative classifier. Two studies in children also confirm the addition of TNF-$\alpha$ to improve distinction between TB patients and healthy individuals (Tebruegge et al., 2015; Tebruegge et al., 2019).

For the read-out of antigen stimulated-blood it has been shown in numerous studies that cytokines other than IFN-$\gamma$ play an important role during the course of infection and may therefore have added diagnostic value (Kassa et al., 2012; Chegou et al., 2012; Belay et al., 2015; Coppola et al., 2016; Tebruegge et al., 2019). A selection of pro- and anti-inflammatory cytokines was therefore included in our study on the basis of previously published research (Walzl et al., 2011; Meier et al., 2018). Our findings suggest that measuring IFN-$\gamma$ only has limited diagnostic potential and that measurement of other cytokines has clear added diagnostic value. In particular, IP-10—a chemokine produced by antigen-presenting cells and induced by a large number of cytokines including IFN-$\alpha$, IFN-$\beta$, IFN-$\gamma$, IL-1$\beta$, IL-2, IL-17, IL-23, TNF-$\alpha$ (Hassanshahi et al., 2007; Mohty et al., 2010)—has been shown to be important in previous studies and our current study. In our study IP-10 concentrations were generally high for all antigens, which were also noted in earlier studies in children (Latorre et al., 2014; Jenum et al., 2016; Petrone et al., 2018). The high measurable concentrations of this cytokine may improve robustness of immunodiagnostic tests especially in children and immunocompromised individuals (Ruhwald et al., 2012). Several studies in adults have shown elevated IP-10 responses in TB disease patients compared to controls (Chegou et al., 2009; Kabeer et al., 2010; Ruhwald et al., 2011). Furthermore antigen-induced IP-10 concentrations were higher in TB disease patients and children from high endemic countries and high-risk groups (Ruhwald et al., 2008; Lighter et al., 2009). One further important aspect particularly interesting for studies in children is the fact that several previous studies suggest IP-10 may be less affected by age as compared to IFN-$\gamma$ (Lighter et al., 2009; Lighter-Fisher et al., 2010). By contrast there are some studies that did find an age-association for IP-10 concentrations (Ruhwald et al., 2008; Decker et al., 2017). Earlier work from our group in healthy children only found an age-association for *Candida albicans*-induced IP-10 concentrations but not for other stimuli (Decker et al., 2017). GM-CSF is thought to have a protective role in the control of TB infection. In our study latency associated antigen Rv2031c induced differential GM-CSF response in healthy and sick individuals. Studies in mice show that deficiency in GM-CSF results in the inability to contain infection (Gonzalez-Juarrero et al., 2005). Other research suggests that survival of bacteria in macrophages is regulated by GM-CSF response in macrophages (Bryson et al., 2019).

In our study we demonstrate the impact of normalization on data with improved performance of a discriminative classifier. Performance was best and most robust when both cytokine-antigen concentrations and between-patient values were normalized. IP-10 concentrations induced by Rv2346/47c and Rv3614/15c were found as major contributors to the performance of the discriminative classifier throughout all normalization methods, likely resulting from high concentrations of this cytokine. However, for cytokines that are not expressed at high concentrations, we show that normalization is highly important. For example, IL-2 and IFN-$\gamma$ concentrations induced by ESAT-6/CFP-10 and Rv3614/15c were only shown to be among the most important features after normalization.

One potential limitation of our study is the sample size which was limited for the two subgroups of TB infection and disease. For optimal training of the classifier and differentiation between TB infection and disease a larger sample size is required. Further studies including a larger number of children are therefore needed to confirm and expand our results. In addition, this study is conducted in a low incidence setting and major factors influencing immune responses such as malnutrition, HIV-infection and other immunocompromising conditions are rare and can therefore not be evaluated.

In conclusion, this is the first study using machine learning algorithms to analyze results from novel *Mycobacterium tuberculosis* antigens and cytokines for the immunodiagnosis of TB in children. The use of machine learning algorithms is a key tool to evaluate large immunological datasets. We identified antigen–cytokine pairs that perform better than the current standard antigen–cytokine pair used in interferon-gamma release assays. These results show that novel antigen–cytokine pairs have to potential to improve immunodiagnostic tests for tuberculosis in children.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethikkommission Nordwestschweiz. Written

informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

## MEMBERS OF THE CITRUS STUDY TEAM

Andrea Duppenthaler, Anne Mornand, Christa Relly, Christian Kahlert, Christoph Berger, Isabelle Rochat Guignard, Jürg Barben, Deborah Levet, Lisa Kottanattu, Marie Rohr, Michael Buettcher, Sara Bernhard-Stirnemann and Nicole Ritz.

## AUTHOR CONTRIBUTIONS

NM and NR developed the research question and the study design. NM performed the experiments. TS, NM, JV, and NR performed the data analysis. NM and NR wrote the draft manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

NM was supported by the following associations: Bangerter Rhyner Stiftung, Lunge Zürich, Nora van Meeuwen-Häfliger Stiftung, Rozalia Foundation, Schweizerische Lungenstiftung and Nikolaus and Bertha Burckhardt Bürgin Foundation.

## ACKNOWLEDGMENTS

The authors thank all the participating centers for their effort in recruiting patients. A special thank goes to Andrea Zelmer for her help with processing samples and the technical support. Also thanks to Kees Franken who purified the *Mycobacterium tuberculosis* antigens. We also like to thank the children and parents for participating in this study.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2020.594030/full#supplementary-material

**SUPPLEMENTARY FIGURE 1 |** relative median cytokine concentrations, (color change from dark blue to light green indicates an increase in relative cytokine concentration). Non-normalized data **(A)**, min–max normalized data **(B)**, mean–std normalized data **(C)**, min–max normalized and mean–std normalized (between individuals) data **(D)**.

**SUPPLEMENTARY FIGURE 2 |** Normalization of data contributes to the performance of a discriminative classifier. Combination of 10 antigen–cytokine pairs contributing the most to the performance of a trained discriminative classifier according to different normalization methods applied: **(A)** non-normalized data **(B)** min–max normalized data and **(C)** mean–std normalized data.

**SUPPLEMENTARY FIGURE 3 |** Normalized cytokine concentrations for individual patients (n = 71) and selected antigen–cytokine pairs sorted by clusters **(A)** and study group **(B)** (color change from dark blue to light green indicates an increase in relative cytokine concentration.

## REFERENCES

Bahk, Y. Y., Kim, S. A., Kim, J. S., Euh, H. J., Bai, G. H., Cho, S. N., et al. (2004). Antigens secreted from Mycobacterium tuberculosis: identification by proteomics approach and test for diagnostic marker. *Proteomics* 4 (11), 3299–3307. doi: 10.1002/pmic.200400980

Belay, M., Legesse, M., Mihret, A., Bekele, Y., Ottenhoff, T. H., Franken, K. L., et al. (2015). Pro- and anti-inflammatory cytokines against Rv2031 are elevated during latent tuberculosis: a study in cohorts of tuberculosis patients, household contacts and community controls in an endemic setting. *PloS One* 10 (4), e0124134. doi: 10.1371/journal.pone.0124134

Bryson, B. D., Rosebrock, T. R., Tafesse, F. G., Itoh, C. Y., Nibasumba, A., Babunovic, G. H., et al. (2019). Heterogeneous GM-CSF signaling in macrophages is associated with control of Mycobacterium tuberculosis. *Nat. Commun.* 10 (1), 2329. doi: 10.1038/s41467-019-10065-8

Chegou, N. N., Black, G. F., Kidd, M., van Helden, P. D., and Walzl, G. (2009). Host markers in QuantiFERON supernatants differentiate active TB from latent TB infection: preliminary report. *BMC Pulm. Med.* 9, 21. doi: 10.1186/1471-2466-9-21

Chegou, N. N., Essone, P. N., Loxton, A. G., Stanley, K., Black, G. F., van der Spuy, G. D., et al. (2012). Potential of host markers produced by infection phase-dependent antigen-stimulated cells for the diagnosis of tuberculosis in a highly endemic area. *PloS One* 7 (6), e38501. doi: 10.1371/annotation/bc36a9c6-d5c0-4d55-bc92-9ce4a07b4f70

Chen, W., Bao, Y., Chen, X., Burton, J., Gong, X., Gu, D., et al. (2016). Mycobacterium tuberculosis PE25/PPE41 protein complex induces activation and maturation of dendritic cells and drives Th2-biased immune responses. *Med. Microbiol. Immunol.* 205 (2), 119–131. doi: 10.1007/s00430-015-0434-x

Commandeur, S., van Meijgaarden, K. E., Lin, M. Y., Franken, K. L., Friggen, A. H., Drijfhout, J. W., et al. (2011). Identification of human T-cell responses to Mycobacterium tuberculosis resuscitation-promoting factors in long-term latently infected individuals. *Clin. Vaccine Immunol.* 18 (4), 676–683. doi: 10.1128/CVI.00492-10

Commandeur, S., van Meijgaarden, K. E., Prins, C., Pichugin, A. V., Dijkman, K., van den Eeden, S. J., et al. (2013). An unbiased genome-wide Mycobacterium tuberculosis gene expression approach to discover antigens targeted by human T cells expressed during pulmonary infection. *J. Immunol.* 190 (4), 1659–1671. doi: 10.4049/jimmunol.1201593

Coppola, M., van Meijgaarden, K. E., Franken, K. L., Commandeur, S., Dolganov, G., Kramnik, I., et al. (2016). New Genome-Wide Algorithm Identifies Novel In-Vivo Expressed Mycobacterium Tuberculosis Antigens Inducing Human T-Cell Responses with Classical and Unconventional Cytokine Profiles. *Sci. Rep.* 6, 37793. doi: 10.1038/srep37793

Decker, M. L., Gotta, V., Wellmann, S., and Ritz, N. (2017). Cytokine profiling in healthy children shows association of age with cytokine concentrations. *Sci. Rep.* 7 (1), 17842. doi: 10.1038/s41598-017-17865-2

Diel, R., Loddenkemper, R., and Nienhaus, A. (2010). Evidence-based comparison of commercial interferon-gamma release assays for detecting active TB: a metaanalysis. *Chest* 137 (4), 952–968. doi: 10.1378/chest.09-2350

Dodge, Y. (2006). *The Oxford Dictionary of Statistical Terms* Ed. Y Dodge (Oxford: Oxford University Press).

Franken, K. L., Hiemstra, H. S., van Meijgaarden, K. E., Subronto, Y., den Hartigh, J., Ottenhoff, T. H., et al. (2000). Purification of his-tagged proteins by immobilized chelate affinity chromatography: the benefits from the use of organic solvent. *Protein Expr. Purif.* 18 (1), 95–99. doi: 10.1006/prep.1999.1162

Goletti, D., Butera, O., Vanini, V., Lauria, F. N., Lange, C., Franken, K. L., et al. (2010). Response to Rv2628 latency antigen associates with cured tuberculosis and remote infection. *Eur. Respir. J.* 36 (1), 135–142. doi: 10.1183/09031936.00140009

Gonzalez-Juarrero, M., Hattle, J. M., Izzo, A., Junqueira-Kipnis, A. P., Shim, T. S., Trapnell, B. C., et al. (2005). Disruption of granulocyte macrophage-colony stimulating factor production in the lungs severely affects the ability of mice to control Mycobacterium tuberculosis infection. *J. Leukoc. Biol.* 77 (6), 914–922. doi: 10.1189/jlb.1204723

Graham, S. M., Cuevas, L. E., Jean-Philippe, P., Browning, R., Casenghi, M., Detjen, A. K., et al. (2015). Clinical Case Definitions for Classification of Intrathoracic Tuberculosis in Children: An Update. *Clin. Infect. Dis.* 61Suppl 3, S179–S187. doi: 10.1093/cid/civ581

Hanley, J. A., and Mcneil, B. J. (1982). The Meaning and Use of the Area Under a Receiver Operating Characterstics (ROC) Curve. *Radiology* 143 (1), 29–36. doi: 10.1148/radiology.143.1.7063747

Hassanshahi, G., Jafarzadeh, A., Ghorashi, Z., Zia Sheikholeslami, N., and Dickson, A. J. (2007). Expression of IP-10 chemokine is regulated by pro-inflammatory cytokines in cultured hepatocytes. *Iran. J. Allergy Asthma Immunol.* 6 (3), 115–121.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12 (1), 55–67. doi: 10.1080/00401706.1970.10488634

Hozumi, H., Tsujimura, K., Yamamura, Y., Seto, S., Uchijima, M., Nagata, T., et al. (2013). Immunogenicity of dormancy-related antigens in individuals infected with Mycobacterium tuberculosis in Japan. *Int. J. Tuberc. Lung Dis.* 17 (6), 818–824. doi: 10.5588/ijtld.12.0695

Jenum, S., Dhanasekaran, S., Ritz, C., Macaden, R., Doherty, T. M., Grewal, H. M. S., et al. (2016). Added Value of IP-10 as a Read-Out of Mycobacterium tuberculosis Specific Immunity in Young Children. *Pediatr. Infect. Dis. J.* 35 (12), 1336–1338. doi: 10.1097/INF.0000000000001328

Kabeer, B. S. A., Raman, B., Thomas, A., Perumal, V., and Raja, A. (2010). Role of QuantiFERON-TB Gold, Interferon Gamma Inducible Protein-10 and Tuberculin Skin Test in Active Tuberculosis Diagnosis. *PloS One* 5 (2). doi: 10.1371/journal.pone.0009051

Kassa, D., Ran, L., Geberemeskel, W., Tebeje, M., Alemu, A., Selase, A., et al. (2012). Analysis of immune responses against a wide range of Mycobacterium tuberculosis antigens in patients with active pulmonary tuberculosis. *Clin. Vaccine Immunol.* 19 (12), 1907–1915. doi: 10.1128/CVI.00482-12

Latorre, I., Díaz, J., Mialdea, I., Serra-Vidal, M., Altet, N., Prat, C., et al. (2014). IP-10 is an accurate biomarker for the diagnosis of tuberculosis in children. *J. Infect.* 69 (6), 590–599. doi: 10.1016/j.jinf.2014.06.013

Leyten, E. M., Lin, M. Y., Franken, K. L., Friggen, A. H., Prins, C., van Meijgaarden, K. E., et al. (2006). Human T-cell responses to 25 novel antigens encoded by genes of the dormancy regulon of Mycobacterium tuberculosis. *Microbes Infect.* 8 (8), 2052–2060. doi: 10.1016/j.micinf.2006.03.018

Li, G., Li, F., Zhao, H. M., Wen, H. L., Li, H. C., Li, C. L., et al. (2017). Evaluation of a New IFN-gamma Release Assay for Rapid Diagnosis of Active Tuberculosis in a High-Incidence Setting. *Front. Cell. Infect. Microbiol.* 7, 117. doi: 10.3389/fcimb.2017.00117

Lighter, J., Rigaud, M., Huie, M., Peng, C. H., and Pollack, H. (2009). Chemokine IP-10: an adjunct marker for latent tuberculosis infection in children. *Int. J. Tuberculosis Lung Dis.* 13 (6), 731–736.

Lighter-Fisher, J., Peng, C. H., and Tse, D. B. (2010). Cytokine responses to QuantiFERON® peptides, purified protein derivative and recombinant ESAT-6 in children with tuberculosis. *Int. J. Tuberculosis Lung Dis.* 14 (12), 1548–1555.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1). (Berkeley, Calif.: University of California Press). p.281–p.297. Available at: https://projecteuclid.org/euclid.bsmsp/1200512992.

Malen, H., Berven, F. S., Fladmark, K. E., and Wiker, H. G. (2007). Comprehensive analysis of exported proteins from Mycobacterium tuberculosis H37Rv. *Proteomics* 7 (10), 1702–1718. doi: 10.1002/pmic.200600853

Mandalakas, A. M., Detjen, A. K., Hesseling, A. C., Benedetti, A., and Menzies, D. (2011). Interferon-gamma release assays and childhood tuberculosis: systematic review and meta-analysis. *Int. J. Tuberc. Lung Dis.* 15 (8), 1018–1032. doi: 10.5588/ijtld.10.0631

Meier, N. R., Jacobsen, M., Ottenhoff, T. H. M., and Ritz, N. (2018). A Systematic Review on Novel Mycobacterium tuberculosis Antigens and Their Discriminatory Potential for the Diagnosis of Latent and Active Tuberculosis. *Front. Immunol.* 9, 2476. doi: 10.3389/fimmu.2018.02476

Mensah, G. I., Addo, K. K., Tetteh, J. A., Sowah, S., Loescher, T., Geldmacher, C., et al. (2014). Cytokine response to selected MTB antigens in Ghanaian TB patients, before and at 2 weeks of anti-TB therapy is characterized by high expression of IFN-gamma and Granzyme B and inter- individual variation. *BMC Infect. Dis.* 14, 495. doi: 10.1186/1471-2334-14-495

Michelsen, S. W., Soborg, B., Diaz, L. J., Hoff, S. T., Agger, E. M., Koch, A., et al. (2017). The dynamics of immune responses to Mycobacterium tuberculosis during different stages of natural infection: A longitudinal study among Greenlanders. *PloS One* 12 (6), e0177906. doi: 10.1371/journal.pone.0177906

Millington, K. A., Fortune, S. M., Low, J., Garces, A., Hingley-Wilson, S. M., Wickremasinghe, M., et al. (2011). Rv3615c is a highly immunodominant RD1 (Region of Difference 1)-dependent secreted antigen specific for Mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci. U. S. A.* 108 (14), 5730–5735. doi: 10.1073/pnas.1015153108

Mohty, A. M., Grob, J. J., Mohty, M., Richard, M. A., Olive, D., and Gaugler, B. (2010). Induction of IP-10/CXCL10 secretion as an immunomodulatory effect of low-dose adjuvant interferon-alpha during treatment of melanoma. *Immunobiology* 215 (2), 113–123. doi: 10.1016/j.imbio.2009.03.008

Oesch Nemeth, G., Nemeth, J., Altpeter, E., and Ritz, N. (2014). Epidemiology of childhood tuberculosis in Switzerland between 1996 and 2011. *Eur. J. Pediatr.* 173 (4), 457–462. doi: 10.1007/s00431-013-2196-z

Perez-Velez, C. M., and Marais, B. J. (2012). Tuberculosis in children. *N. Engl. J. Med.* 367 (4), 348–361. doi: 10.1056/NEJMra1008049

Petrone, L., Vanini, V., Chiacchio, T., Petruccioli, E., Cuzzi, G., Schinina, V., et al. (2018). Evaluation of IP-10 in Quantiferon-Plus as biomarker for the diagnosis of latent tuberculosis infection. *Tuberculosis* 111, 147–153. doi: 10.1016/j.tube.2018.06.005

Ruhwald, M., Petersen, J., Kofoed, K., Nakaoka, H., Cuevas, L. E., Lawson, L., et al. (2008). Improving T-Cell Assays for the Diagnosis of Latent TB Infection: Potential of a Diagnostic Test Based on IP-10. *PloS One* 3 (8). doi: 10.1371/journal.pone.0002858

Ruhwald, M., Dominguez, J., Latorre, I., Losi, M., Richeldi, L., Pasticci, M. B., et al. (2011). A multicentre evaluation of the accuracy and performance of IP-10 for the diagnosis of infection with M. tuberculosis. *Tuberculosis* 91 (3), 260–267. doi: 10.1016/j.tube.2011.01.001

Ruhwald, M., Aabye, M. G., and Ravn, P. (2012). IP-10 release assays in the diagnosis of tuberculosis infection: current status and future directions. *Expert Rev. Mol. Diagn.* 12 (2), 175–187. doi: 10.1586/erm.11.97

Serra-Vidal, M. M., Latorre, I., Franken, K. L., Diaz, J., de Souza-Galvao, M. L., Casas, I., et al. (2014). Immunogenicity of 60 novel latency-related antigens of Mycobacterium tuberculosis. *Front. Microbiol.* 5, 517. doi: 10.3389/fmicb.2014.00517

Sollai, S., Galli, L., de Martino, M., and Chiappini, E. (2014). Systematic review and meta-analysis on the utility of Interferon-gamma release assays for the diagnosis of Mycobacterium tuberculosis infection in children: a 2013 update. *BMC Infect. Dis.* 14. doi: 10.1186/1471-2334-14-S1-S6

Tebruegge, M., Dutta, B., Donath, S., Ritz, N., Forbes, B., Camacho-Badilla, K., et al. (2015). Mycobacteria-Specific Cytokine Responses Detect Tuberculosis Infection and Distinguish Latent from Active Tuberculosis. *Am. J. Respir. Crit. Care Med.* 192 (4), 485–499. doi: 10.1164/rccm.201501-0059OC

Tebruegge, M., Ritz, N., Donath, S., Dutta, B., Forbes, B., Clifford, V., et al. (2019). Mycobacteria-Specific Mono- and Polyfunctional CD4+ T Cell Profiles in Children With Latent and Active Tuberculosis: A Prospective Proof-of-Concept Study. *Front. Immunol.* 10, 431. doi: 10.3389/fimmu.2019.00431

Tundup, S., Mohareer, K., and Hasnain, S. E. (2014). Mycobacterium tuberculosis PE25/PPE41 protein complex induces necrosis in macrophages: Role in virulence and disease reactivation? *FEBS Open Bio* 4, 822–828. doi: 10.1016/j.fob.2014.09.001

Voskuil, M. I., Schnappinger, D., Visconti, K. C., Harrell, M. I., Dolganov, G. M., Sherman, D. R., et al. (2003). Inhibition of respiration by nitric oxide induces a Mycobacterium tuberculosis dormancy program. *J. Exp. Med.* 198 (5), 705–713. doi: 10.1084/jem.20030205

Walzl, G., Ronacher, K., Hanekom, W., Scriba, T. J., and Zumla, A. (2011). Immunological biomarkers of tuberculosis. *Nat. Rev. Immunol.* 11 (5), 343–354. doi: 10.1038/nri2960

World Health Organization (2013). *Roadmap for childhood tuberculosis*. Geneva; WHO.

World Health Organization (2018a). *Global Tuberculosis Report 2018*. Geneva; WHO.

World Health Organization (2018b). *Roadmap towards ending TB in children and adolescents*. Geneva; WHO.

Yao, J., Du, X., Chen, S., Shao, Y., Deng, K., Jiang, M., et al. (2018). Rv2346c enhances mycobacterial survival within macrophages by inhibiting TNF-alpha and IL-6 production via the p38/miRNA/NF-kappaB pathway. *Emerg. Microbes Infect.* 7 (1), 158. doi: 10.1038/s41426-018-0162-6

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2021 Meier, Sutter, Jacobsen, Ottenhoff, Vogt and Ritz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Public Health

ORIGINAL RESEARCH
published: 21 January 2021
doi: 10.3389/fpubh.2020.583401

# Adaptive Time-Dependent Priors and Bayesian Inference to Evaluate SARS-CoV-2 Public Health Measures Validated on 31 Countries

*Hugues Turbé [1,2]*, Mina Bjelogrlic [1,2], Arnaud Robert [1,2], Christophe Gaudet-Blavignac [1,2], Jean-Philippe Goldman [1,2] and Christian Lovis [1,2]*

[1] Medical Information Sciences Division, Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland, [2] Medical Information Sciences Division, Diagnostic Department, University Hospitals of Geneva, Geneva, Switzerland

With the rapid spread of the SARS-CoV-2 virus since the end of 2019, public health confinement measures to contain the propagation of the pandemic have been implemented. Our method to estimate the reproduction number using Bayesian inference with time-dependent priors enhances previous approaches by considering a dynamic prior continuously updated as restrictive measures and comportments within the society evolve. In addition, to allow direct comparison between reproduction number and introduction of public health measures in a specific country, the infection dates are inferred from daily confirmed cases and confirmed death. The evolution of this reproduction number in combination with the stringency index is analyzed on 31 European countries. We show that most countries required tough state interventions with a stringency index equal to 79.6 out of 100 to reduce their reproduction number below one and control the progression of the pandemic. In addition, we show a direct correlation between the time taken to introduce restrictive measures and the time required to contain the spread of the pandemic with a median time of 8 days. This analysis is validated by comparing the excess deaths and the time taken to implement restrictive measures. Our analysis reinforces the importance of having a fast response with a coherent and comprehensive set of confinement measures to control the pandemic. Only restrictions or combinations of those have shown to effectively control the pandemic.

**Keywords: infectious diseases, reproductive number estimation, non-pharmaceutical interventions, Bayesian inference (BI), health sciences, epidemiology, SARS -CoV-2, public health**

OPEN ACCESS

***Edited by:***
*Adrian Egli, University Hospital of Basel, Switzerland*

***Reviewed by:***
*David Alfredo Medina Ortiz, University of Chile, Chile*
*Sebastian Contreras, Max Planck Society (MPG), Germany*

****Correspondence:***
*Hugues Turbé*
*Hugues.turbe@unige.ch*

***Specialty section:***
*This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health*

***Received:*** *22 July 2020*
***Accepted:*** *03 December 2020*
***Published:*** *21 January 2021*

***Citation:***
*Turbé H, Bjelogrlic M, Robert A, Gaudet-Blavignac C, Goldman J-P and Lovis C (2021) Adaptive Time-Dependent Priors and Bayesian Inference to Evaluate SARS-CoV-2 Public Health Measures Validated on 31 Countries. Front. Public Health 8:583401. doi: 10.3389/fpubh.2020.583401*

## INTRODUCTION

Since being first observed in Wuhan in late 2019, the outbreak of the 2019 SARS-CoV-2 virus is strongly affecting societies and economies. The transmission rate, pressure on the healthcare system and lack of effective treatment lead countries to take public health measures to limit the spread of the virus. The confinement measures range from banning gatherings to complete lockdowns and closing borders (1, 2). Additional measures include individual protection with various levels of mask wearing injunctions, and contact tracing with quarantine. This work has focused on developing reliable modeling approaches to evaluate the impact of public health

measures. Our method is based on analyzing the reporting of European countries to evaluate the temporal influence of non-pharmaceutical interventions (NPIs) on the effective reproduction number $R_t$. The aim of $R_t$ is to quantify the number of secondary infections caused by an individual over the period during which this person is infectious. It is important to make the distinction between the effective and basic reproduction number. The basic reproduction number $R_0$ refers to the evolution of the disease when the population is fully susceptible to the disease while $R_t$ factors numerous parameters, such as the susceptible population, the transmission, the public awareness, the immunity acquired within the population, amongst others (3). $R_t$ is a key parameter to evaluate the evolution of an epidemic. Any value below one indicates that the spread is decreasing, any value above one indicates that the spread is increasing in a given population. $R_t$ allows a direct comparison of the epidemiologic profiles observed in different cohorts of population, such as specific risk factors driven cohorts or countries with distinct characteristics (such as population or testing methods). It allows thus to consider temporality and populational or cohorts characteristics. The spatial and populational (age, social activity) heterogeneity have been shown to play a role in the evolution of the pandemic as the $R_t$ evolve differently across these different groups (4–6).

Numerous methods have been developed to compute $R_t$ and its evolution over time (7) with the aim of identifying the most influential parameters and predicting the development of an epidemic in a given environment. Initial methods derived $R_t$ from transmission model similar to the SIR model (8–12). In general, fitting deterministic model to incidence data has been shown to often results in large error which can however be solved by using stochastic model (13). The choice of the mechanistic transmission model requires assumptions about the epidemiology of the disease. For example, the presence/absence of a latency period will guide the choice between a SIR (Susceptible—Infected—Recovered) or SEIR (Susceptible—Exposed—Infected—Recovered) model. Recent studies tend to acknowledge the risk of asymptomatic transmission of COVID-19 although with a lower relative risk than transmission by symptomatic individuals (14) favoring the use of a SIR model. The latter model is parametrized through the transmission rate $\beta$ and the rate of removal $\gamma$. One pitfall is that this model assumes a constant transmission rate, that is the infection probability distribution is constant over the period during which an individual is infectious. In addition, the SIR model requires to be fitted to the number of infections as well as the number of people either susceptible or who have recovered. However, the latter two variables, susceptible and recovered, are difficult to evaluate and will strongly be influenced by underreporting. Later models, including the Wallinga and Teunis approach (15), use a likelihood-based estimation procedure to reconstruct infection patterns. These methods have shown large variations when using daily data (16). Most approaches aiming at correcting these fluctuations appeared to be sensitive to smoothing parameters (16, 17). An additional method to mitigate these drawbacks that is very robust to underreporting was later developed (18). This method used Bayesian inference based on a transmission model which includes the infectivity profile to update the posterior distribution of $R_t$ as more data become available.

Since the start of the COVID-19 pandemic, various studies have looked at the impact of public health interventions on the evolution of the pandemic at regional or national level. The first studies, on data from China, proving the impact of NPI strategies to reduce $R_t$ used mechanistic transmission models to obtain $R_t$ (19, 20), with the drawbacks described above associated with these models. Further studies focused on how state interventions prevented ICU capacity to be overwhelmed as well as their impact on fatalities in the UK (21), Germany (22, 23), and France (24). While these researches focused on individual country, a recent study aimed to demonstrate the impact of non-pharmaceutical interventions in 11 European countries (25). This study assumed that the impact of the measures was independent of their relative introduction. In addition, this study assumed $R_t$ to be fixed between the different measures. However, a recent research shows that community changes also play a role in slowing the evolution of the virus (26).

When evaluating the impact of public health interventions, it is crucial to consider that there is a delay between the time of infection and the time at which a confirmed case or the death of an individual is reported. Even if we consider that NPIs have a direct impact on the rate of infections, there will be a delay between this change of infections and the time at which this change is observed through positives tests or the death of the individuals. The simplest method would consist in shifting the data backward in time by the mean of the distribution of interest that is the period from infections to the case being reported or the death of the individual. However, this method does not account for the uncertainty in the period of interest. A possible method to circumvent this issue consists in subtracting samples from the delay distribution to each observation. This method has been recently used to adjust reporting delays in the aim of evaluating the reproduction number of SARS-CoV-2 (25, 27) and was applied in our research. One drawback of the method is that as the mean and variance of the delay distribution increase, the resulting infections are smoothed over time potentially blurring discontinuities in the variation of $R_t$ (28). Alternatively, the confirmed cases can be considered as the convolution of the infections with a delay period distribution. The process to obtain the time of infection can therefore be performed using a maximum-likelihood deconvolution method (29, 30). These methods build on techniques which were initially develop to correct AIDS data based on an iterative EM algorithm (31). A different approach aimed to jointly infer the infections and $R_t$ (32). The drawback of this method is that it requires an hypothesis on the shape and change points of $R_t$.

The aim of this work is to extend previous research estimating $R_t$ and focusses on the effects of state interventions in 31 European countries. As the evolution of $R_t$ is a function of at least three important parameters: the type of the restrictive measures; the effect of these measures and changes in behaviors with specific societal properties, and the size of various compartmental cohorts involved, we do not aim to quantify the effect of each measure. The restrictive measures and their effects are first considered to be independent across the different countries. We

then compare their effects across the countries and aim to show how the combined interventions within a country and their temporality have influenced the spread of the virus, characterized by the evolution of confirmed cases, confirmed deaths, and excess deaths.

# MATERIALS AND METHODS

The following section aims to describe the different steps of the analysis. The various data sources used in the analysis as well as required period distributions for SARS-CoV-2 are first introduced. Secondly, statistical methods to estimate $R_t$ are formulated and lastly the method to evaluate the impact of NPIs is described.

## Data Sources and Availability

$R_t$ is estimated using incidence data for confirmed cases and deaths published in the *COVID-19 Data Repository* (33).

The excess mortality was retrieved from Our World in Data, (34). The data are aggregated on a weekly basis along the average deaths observed for the same period between 2015 and 2019.

Data related to the period between a positive test and the death of an individual were retrieved from: Swiss Federal Office of Public Health (FOPH) (35). Data from FOPH on confirmed cases is used to evaluate the impact of different information sources.

Data regarding the various state interventions were retrieved from the *Coronavirus government response tracker* (OxCGRT) developed by the Blavatnik School of Government (36). The stringency index provided in this dataset tracks government's policies and interventions across different categories and provides a score between 0 and 100 evaluating the overall stringency of the measures taken in a given country (37). A stringency index of zero means no measure has been noted in this country, and a maximum score of 100, indicates a complete lock down. The stringency index is calculated as averages of the individual component indicators categorized in the following six groups: school closing, non-essential economic activities, public events, gatherings, stay at home policies, and restrictions on movements. For the "Stringency index" the sub-index score $I_{j,t}$ is calculated for the 9 indicators as follows:

$$I_{j,t} = 100\frac{v_{j,t} - 0.5(F_j - f_{j,t})}{N_j} \tag{1}$$

With $N_j$ being the maximum value of the indicator, $F_j$ the indicator flag (whether the measure has or not a sectoral scope), $v_{j,t}$ the recorded policy on the ordinal scale, and finally $f_{j,t}$, being the recorded binary flag for that indicator. The full methodology, the variable values for computing the different scores are available on their github repository, along with the interpretation of each indicator (see Data Availability Statement for the exact reference). The evolution of the stringency index for the countries of interest can be found in **Supplementary Figure 1**.

A dataset which included the intersection of the data regarding the evolution of the confirmed cases and deaths as well as the data measuring the stringency index was available for 33 European countries. For our analysis, Russia and Ukraine were removed from our dataset as the reported daily deaths were still increasing for these two countries when we are interested in countries which have successfully contained the evolution of the pandemic before the 23rd of May 2020. We were therefore left with a set of 31 European countries. The full list of the countries included in the analysis is presented in the results sections. For the second part of the analysis which focused on the excess deaths observed in each country, the data were available for 19 countries.

## Determining Incubation Time, Onset to Confirmed, and Onset to Death Distributions

The proposed method allows to compute $R_t$ without developing a transmission model and hence only requires a hypothesis on the infectivity profile or serial interval distribution. The infectivity profile is a probability distribution measuring the probability to infect an individual at a given time $s$ after the infection of the primary case. This distribution is crucial to model the dynamic of the infections and the delay between the primary and secondary cases. The incidence on a given day can be estimated as follows:

$$E[I_t] = R_t \sum_{s=1}^{t} w_s I_{t-s} \tag{2}$$

where $E[\bullet]$ is the expected value of a random variable, $I_t$ is the incidence at time $t$, and $w_s$ is the infectivity profile. The distribution of $w_s$ for the SARS-CoV-2 virus was found to have a mean of 4.8 days and a standard deviation of 2.3 days (38).

Given the time at which the infection occurred is not available, the number of confirmed cases and deaths on a given day are used as proxies. A gamma distribution with a median incubation period at 4.4 days from confirmed infection and diagnosis outside of the epicenter of Hubei Province, China, based on official reports from governmental institutes was derived (39). The mean and deviation were then obtained by fitting a gamma distribution to the quantile derived in this study. The period between the onset of the symptoms and a case being confirmed in Switzerland, was estimated to 5.6 days (40).

The period between a case being reported as positive and the death of the individual was extracted from 1,430 cases provided by the Swiss Federal Office of Public Health (FOPH). Our result provides a distribution on a much larger dataset than the one built which used between 24 and 33 cases (39, 41). Three different distributions were tested: lognormal, Weibull and gamma with the Akaike Information Criterion (AIC) being used to identify the best distribution. This distribution was then combined with the incubation period (39) to obtain the period between onset and death shown in **Table 1** along the other distribution periods where the onset refers to the symptom onset.

From the latter period functions it is possible to calculate a posterior distribution of $R_t$ based on the inferred infection dates extracted from the confirmed cases and deaths reported. For the daily cases declared (incidence), a shift following a gamma distribution between the defined cases (confirmed or dead) and the time of infection is randomly generated. For each case, the new date of infection is generated by subtracting the shift to the reported date. This procedure is performed iteratively with

**TABLE 1 |** Incubation, onset to confirmed and onset to death distributions where onset refers to symptoms onset.

| Period | Mean [days] | Standard deviation [days] |
|---|---|---|
| Incubation (39) | 4.6 | 1.9 |
| Onset to confirmed (40) | 5.6 | 4.2 |
| Onset to death (our study) | 15.3 | 8.0 |

the mean of daily simulated number of infections stored. Using the latter period functions to estimate the infection occurrences allows to take into account the large variance in the cases reported by the health or political systems in the analyzed countries.

## Correcting the Number of Infections

In addition, the incidence for the most recent days are corrected (40) to factor delayed reporting:

$$\bar{I}_t = \frac{I_t}{\hat{F}_l} \tag{3}$$

where $\bar{I}_t$ and $I_t$ are, respectively, the corrected and initial incidence which took place on a given day. $\hat{F}$ is the cumulative distributive function of the period between an infection and a case being reported as positive or dead, $l$ is the time between $t$ and the last reported case so that $\hat{F}_l = \mathrm{P}(X \leq l)$ where $X$ is a random variable that is gamma distributed with mean and standard deviation described in **Table 1** depending on the variable of interest and $P(X \leq l)$ is the probability that $X$ is smaller or equal to $l$.

## Estimation of the Reproduction Number Using Bayesian Inference With Time-Dependent Priors

The method presented in this report is a variation of the one proposed by Cori et al. (18). Assuming the incidence at time $t$, $I_t$, is Poisson distributed so that the likelihood of the incidence $I_t$ given $R_t$ and conditional on previous incidences $I_0, \cdots, I_{t-1}$:

$$P\left(I_t | I_0, \cdots, I_{t-1},\ w, R_t\right) = \frac{\left(R_t \Lambda_t\right)^{I_t} e^{-R_t \Lambda_t}}{I_t!} \tag{4}$$

with $\Lambda_t = \sum_{s=1}^{t} w_s I_{t-s}$ where $w_s$ is the estimated infectivity profile.

The posterior of $R_t$ conditional on previous incidences is:

$$P\left(R_t | I_0, \cdots, I_{t-1,}\ I_t,\ w\right) \propto P\left(I_t | I_0, \cdots, I_{t-1},\ w, R_t\right) P\left(R_t\right) \tag{5}$$

While the method developed by Cori et al. (18) assumes a constant gamma distribution for the prior distribution, the presented model takes advantage of the information gained in time by updating the prior distribution for each window with the previous posterior:

$$P\left(R_t\right) = P\left(R_{t-1} | I_0, \cdots, I_{t-2}, I_{t-1},\ w\right) \tag{6}$$

The 95% CI is then derived by computing the 2.5% and 97.5% quantiles.

$R_t$ based on the confirmed cases is reported up to 9 days before the last date at which results are available. This corresponds to the median time for confirmed cases to be reported. Using the same method, $R_t$ based on the cases reported as dead is reported up to 19 days before the last day on which deaths were reported for a given country.

## Comparison of the Methods to Estimate $R_t$ on Synthetic Data

In order to compare the proposed methods with the one developed by Cori et al. (18), a study on synthetic data was performed. Two scenarios which were initially used in the aforementioned research were used:

1. Constant reproduction number, $R_t = 2.5$
2. Sharp change in the reproduction number:
   - $R_t = \begin{cases} 2.5\,, & t \leq 15\ days \\ 0.8, & t > 15\ days \end{cases}$

For each scenario, 100 simulations were performed. Ten cases were introduced at $t = 0$ *days*, with the incident cases $I_t$ for the following 49 days being drawn from a Poisson distribution with mean equal to $R_t \sum_{s=1}^{t} I_{t-s}\, w_s$. An infectivity profile $w_s$ with a mean of 4.8 and standard deviation of 2.3 days as introduced by Nishiura et al. (38) for the SARS-CoV-2 virus was used. $R_t$ was then evaluated from the synthetic data using the method developed by Cori et al. (18) as well as the proposed method.

The impact of underreporting was simulated using a binomial distribution as performed in (18). For each day, the new incident cases $I_t^*$ were assumed to follow a binomial distribution:

$$I_t^* \sim Binomial(I_t,\ \pi) \tag{7}$$

where $\pi$ is the reporting rate and was varied between 20 and 80% in steps of 20%. $R_t$ was then evaluated on the simulated underreported data and compared to the simulated $R_t$.

## Assessing NPIs' Impact on the Evolution of the Pandemic

The stringency index developed as part of the OxCGRT project (37) was used to assess the role of state interventions in controlling the pandemic. This index was compared with the evolution of $R_t$, rather than the incidence of confirmed or dead cases. Using $R_t$ helps comparing countries that have heterogeneous testing or reporting policies. While $R_t$ is also subject to variations in these policies, it depends on the change within the country in confirmed and death cases, therefore allowing comparison between countries with different policies. For each country, the public health measures and the stringency index are analyzed when $R_t$ estimates, based on the confirmed cases, dropped below one. The hypothesis is that it can help identifying the most efficient set of public health measures.

In order to assess the impact of taking restrictive measures early in the crisis, the time taken to introduce initial restrictive measures was compared to the period taken to control the

epidemic. The time until the introduction of restrictive measure was defined as the period between the 5th death in a given country and the stringency index reaching a score of 35. The stringency index threshold at 35 corresponds to the lowest score observed when a country reached a $R_t$ smaller than one which was observed for Andorra. The time required to control the epidemic was then defined as the period between the 5th death and $R_t$, based on the confirmed cases, dropping below one.

Given that the confirmed cases and reported deaths are influenced by reporting policies, the analysis described above was supported by using the number of excess deaths. Following the same logic as for the previous analysis, the period between the 5th death and the stringency index reaching 35 was compared to the excess deaths experienced in each country. The excess deaths were calculated as:

$$Excess\ deaths = \sum_w \frac{Deaths|_{Week\#w\ 2020} - Average\ Deaths|_{Week\#w\ 2015-2019}}{Average\ Deaths|_{Week\#w\ 2015-2019}} \tag{8}$$

where $w$ represents for each country the weeks between the 5th death and the 23rd of May 2020. This alternative method to measure the impact of the different NPIs independently of the proposed method to compute $R_t$ serves as a mean to support our conclusions.

# RESULTS

## Evaluation of the Proposed Methods

The simulated incident cases described in section Comparison of the methods to estimate $R_t$ on synthetic data are presented in **Figure 1** for the two scenarios used to validate the proposed methods. The $R_t$ computed using the proposed method as well as the one from (18) for the first scenario are shown in **Supplementary Figure 2**, while the results for the second scenario which includes a discontinuity in the simulated $R_t$ are shown in **Figure 2.** In order to compare the two methods, the average relative error was computed using the following equation:

$$Error = \frac{1}{l} \sum_{t=0}^{l} \frac{|R_t - \overline{R_t}|}{\overline{R_t}} \tag{9}$$

where l is the number of days for which the computed $R_t$ can be derived from the simulated incident cases and $\overline{R_t}$ is the simulated reproduction number over the same period $l$.

The computed average relative errors for the two scenarios and methods are presented in **Table 2.**

The $R_t$ evaluated on the simulated underreported data following the method described in section Comparison of the methods to estimate $R_t$ on synthetic data are presented for a reporting rate of 20, 40, 60, and 80% for the two scenarios in **Supplementary Figures 3–8**. The average relative errors for these simulations are shown in **Table 3.**

The proposed method takes as input confirmed cases which can be provided by different sources (health or political systems). In **Supplementary Figure 9**, the reproduction number is estimated for Switzerland, with two different sources.

## Evaluating the Reproduction Number From Incidence Data of 31 Countries

The list of countries analyzed along dates characterizing the evolution of the epidemic and stringency index values are listed in **Table 4** which is composed of four panels. This table summarizes our analysis performed by computing $R_t$, based on the confirmed cases. The first panel includes the dates which were used to characterize the evolution of the pandemic in each country. The first column of this panel is the date at which the 5th death was observed, the 2nd one when the stringency index reached a value of 35 and the third one includes the date at which the country managed to control the epidemic by reducing $R_t$, below one. The second panel shows the value of the stringency index when $R_t$, was reduced below one. The third shows the period between the 5th death and the stringency index reaching 35 or $R_t$ becoming smaller than one. The last panel includes the computed excess deaths. The same table with the data when $R_t$ is evaluated on the reported deaths can be found in **Supplementary Table 1**.

As a case study, the evolution of $R_t$ in Austria is shown in **Figure 3**. **Figure 3** aims to illustrate the different steps of the analysis and will be used for the discussion. In the top part, the daily confirmed cases are shown as a histogram. From these daily confirmed cases and the derived period distributions, the inferred daily infection are displayed as a dashed line. In the middle part, the mean estimated $R_t$ is displayed as a full line, along with its 95% CI as a shaded area, with $R_t$ being estimated from the inferred infections. In the bottom part, the evolution of the stringency index is displayed with a colorbar changing toward dark red as the stringency score goes toward its maximum value of 100, through the period of interest (from the date of the 5th death up to the 23rd of May). Different interesting phases of the pandemic are shown in the Austrian example depicted in **Figure 3**. Firstly, $R_t$ started to decline before the introduction of restrictive measures between March 13th and 17th, and this reduction was intensified by a combination of NPIs which sums into a high stringency index score. $R_t$ then plateaued at around 0.65 during the lockdown and has been oscillating around one up to the end date of our analysis (23rd of May). This last phase shows the emergence of localized clusters.

When countries managed to reduce their $R_t$ estimated on the confirmed cases below one, they had a mean stringency index of 79.6 out of 100 with a standard deviation of 14.3. The individual stringency indices for each country are presented in **Figure 4.** When $R_t$ dropped below one, the median severity of the measures along their individual severity out of 100 for each category defined in the OxCGRT dataset was the following: (a) School closed (100/100); (b) Non-essential economic activities closed (100/100); (c) Public events were canceled (100/100); (d) Gathering of more than 10 people banned (100/100); (e) Mandatory at home policy with minimal exceptions (67/100); (f) Movements in the country were restricted (100/100). **Figure 5** shows the time from the 5th death to $R_t$ reducing below one against the time from

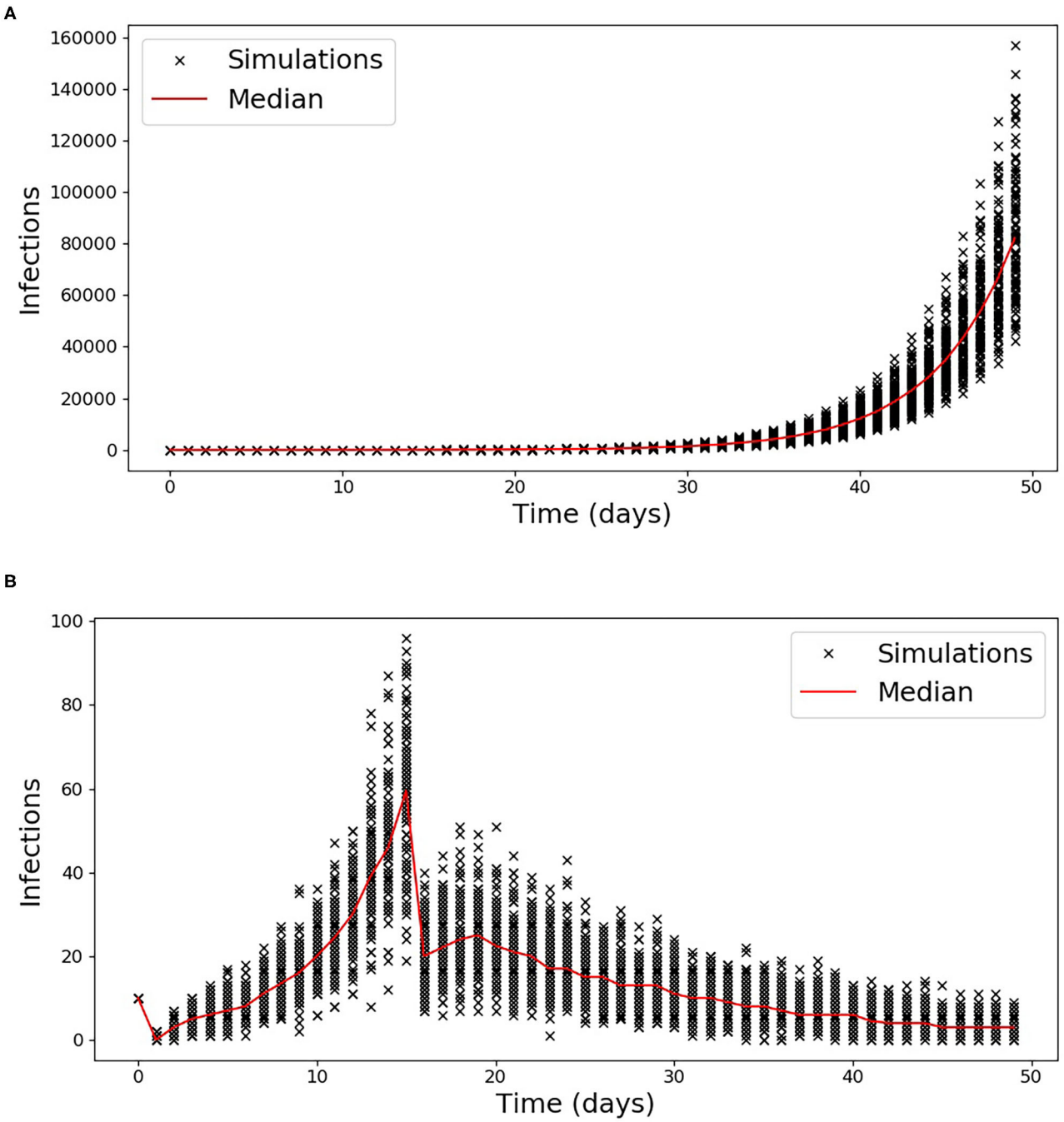



**FIGURE 1 |** Simulated Incident cases for scenario 1 **(A)** and 2 **(B)**. The incident cases for each of the 100 simulations are reported with a cross while the median is indicated by the red line.

the 5th death to the date where the stringency reached 35. A Pearson correlation coefficient of 0.722 was found between the two variables.

This analysis was repeated for the $R_t$ estimated on the reported deaths. A gamma distribution with a mean and a standard deviation equal, respectively, to 9.7 and 6.73 days was found, using the AIC criterion, to best fit the data from a case being confirmed to its death. The distribution along the extracted data are shown in **Supplementary Figure 10**. The AIC for the different distribution are summarized in **Supplementary Table 2**. This distribution was used to estimate the $R_t$ on the confirmed deaths. A Pearson correlation coefficient of 0.338 was obtained between the two variables, that is the time between the 5th death and the stringency index reaching

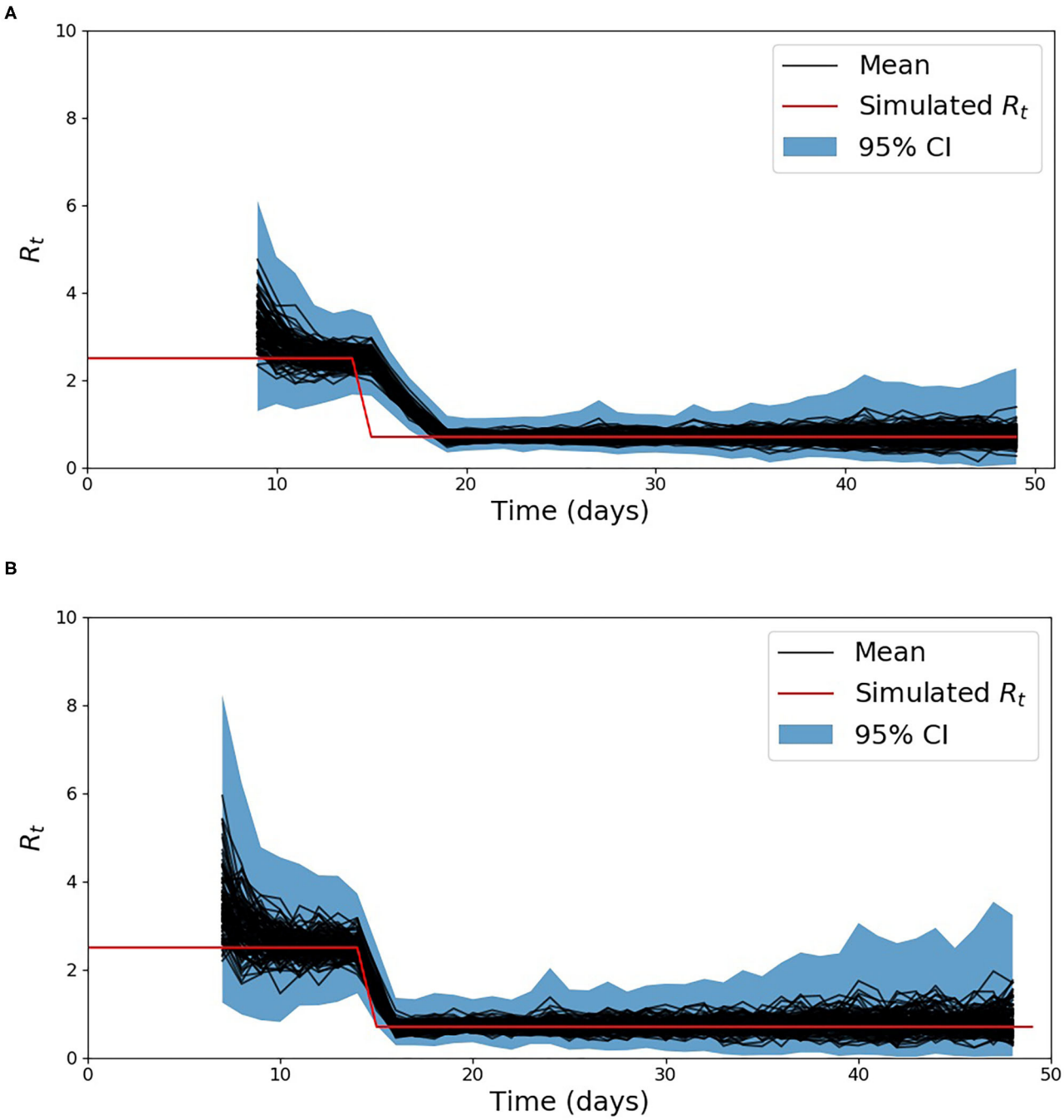



**FIGURE 2 |** Estimation of $R_t$ for scenario 2, using the baseline method developed by Cori et al. (18) **(A)** and the developed method **(B)**.

35 and the time between the 5th death and the $R_t$ reducing below one. The results for this analysis are presented in **Supplementary Figures 11**, **12**.

The comparison between the level of excess deaths observed in a given country and the time between the 5$^{th}$ death and the stringency index reaching 35 are presented in **Figure 6**. A Pearson correlation of 0.684 was observed between these two variables.

# DISCUSSION

## Evaluation of the Proposed Method to Estimate $R_t$

The method developed to estimate the effective reproduction number $R_t$ is based on the method developed by Cori et al. (18). This method only requires the infectivity profile and an initial assumptions of the basic reproduction number $R_0$

**TABLE 2 |** Average relative error comparison between the proposed method and the one developed by Cori et al. (18) measured on synthetic data.

| | Average relative error [%] | | |
|---|---|---|---|
| | **Baseline method** | **Proposed method** | **Δ [%]** |
| Scenario 1 | 1.81 | 0.87 | −51.9 |
| Scenario 2 | 17.7 | 9.01 | −49.2 |

**TABLE 3 |** Average relative error comparison between the proposed method and the one developed by Cori et al. (18) measured on the underreported synthetic data.

| | | Average relative error [%] | | |
|---|---|---|---|---|
| **Underreporting rate $\pi$** | **Scenario** | **Baseline method** | **Proposed method** | **Δ [%]** |
| 0.2 | 1 | 5.34 | 5.3 | −0.75 |
| | 2 | 28.91 | 31.81 | 10.03 |
| 0.4 | 1 | 2.27 | 1.72 | −24.23 |
| | 2 | 20.02 | 17.14 | −14.39 |
| 0.6 | 1 | 2.06 | 1.23 | −40.29 |
| | 2 | 19.3 | 12.1 | −37.31 |
| 0.8 | 1 | 1.8 | 1.04 | −42.22 |
| | 2 | 18.4 | 9.78 | −46.85 |

used to initialize the prior. The difference and main advantage of the proposed method is that we are less reliant on the initial assumptions of $R_0$. While (18) assumes the prior is fixed in time, we constantly adapt it with new data. As seen in **Supplementary Figure 2**, for a constant reproduction number both methods, the baseline and the proposed method converge toward the simulated value of 2.5. The similarity between the two methods on this scenario is also reflected in the average relative error presented in **Table 2**. Both methods have a low error, but the proposed method reduces the average error by around 1%. This reduction is mainly due to its faster rate of convergence toward the start of the simulated data. The difference between the two methods are more visible in the 2nd scenario which simulate a discontinuity in $R_t$. This discontinuity aimed to simulate the extreme case where the introduction of a given NPI would have a direct effect on $R_t$. As seen in **Figure 2**, the developed method tracks the sharp change in $R_t$ arising on day 15 much more closely than the baseline method. As a result the average relative error over the simulations reduces from 17.7% with the baseline method to 9.0% with the developed method. This result is expected given that the distribution's prior is updated with the most recent data, while in the method proposed by Cori et al. (18), only the posterior evolves.

The baseline method was shown by its authors to be robust to underreporting (18). Given it is a known issue in the current pandemic and it was even more so toward the start of the pandemic, it was important to verify than the proposed methods retained this beneficial characteristic. As described in section Comparison of the methods to estimate $R_t$ on synthetic data, underreporting was simulated on the synthetic data. As shown in **Table 3,** the developed method overperformed the baseline one in all simulated cases, except the 2nd scenario with a reporting factor of 20%. This error mainly arises from the incident cases which lies at the end of the simulated periods with only one or two incident cases being simulated over the last 15 days. Over all simulations which replicates underreporting (**Supplementary Figures 3–8**), the proposed method has a larger confidence interval when $R_t$ is estimated on very small incident cases.

## Challenges in Estimating $R_t$ on Real Data

As it is very difficult at the beginning of an epidemic to correctly evaluate $R_0$ (42), it is important to update the prior as more data become available. In the future, our method will therefore be generalizable to new epidemic and provide reliable data at the start of the epidemic by being less reliant on the initial estimation of $R_0$. However, as previous methods developed to estimate $R_t$, our method is sensible to change in testing policy within a given country. It is also important to note that as there is a delay between the infection of an individual and the individual testing positive or dying, the $R_t$ measured today reflects the evolution of the pandemic shifted in the past by the distribution of the period between the infection and the case being confirmed or the death of the individual. Models aiming to correct this delay have been initially developed to correct the data following the delay between a positive test and the test being reported (43) in order to allow real-time tracking of epidemics. More recently, Nowcasting methods using hierarchical Bayesian model have been used to provide reliable and up-to-date estimate of the $R_t$ (44).

Confirmed cases and deaths are widely available in the public domain, but to estimate the infection dates, the incubation period and the period between the onset of the symptoms and the person having a positive test or the death of the individual is required. The incubation period was initially derived on Chinese cases (39) and it was assumed that this property is intrinsic to the virus and is therefore relevant for European countries. The period between the symptoms onset and a case being confirmed has been derived on Swiss patient (40). The period between the symptoms onset and the death of the patient was derived on Chinese data (39), but this period was not available for European patients. Based on 1,430 Swiss cases, we found this period to have a mean of 15.3 days compared to 16.3 days in Linton et al. (39). It was then assumed that this period was relevant for the European countries included in our study. All the periods distribution used for the rest of the analysis are summarized in **Table 1**.

Impact of data sources have been qualitatively evaluated for Switzerland. $R_t$ has been separately estimated on data from the international repository of JHU (33) and the national repository of FOPH (35) for the same period of time (**Supplementary Figure 9**). The average relative error equation (9) between the two estimated $R_t$ is 6%. This value is relatively low compared to the changes in the reported cases. As an example, $R_t$ dropped below one for the first time for both estimates on the 18th of March, even though on the exact same day, the confirmed new cases were reported to be, respectively, 328 and 1,211, for JHU and FOPH sources. As visible in

**TABLE 4 |** List of countries along dates characterizing the evolution of the epidemic (with $R_t$ measured on the confirmed cases) and measured excess deaths in percent of the number of average death observed between 2015 and 2019.

| | Date | | | Value | Days from 5th death to: | | |
|---|---|---|---|---|---|---|---|
| | 5th death | $Str_{idx} > 35$ | $R_{t<1}$ | $Str_{idx}$ when $R_{t<1}$ | $Str_{idx} > 35$ | $R_{t<1}$ | Excess death [%] |
| Albania | 26.03.2020 | 09.03.2020 | 31.03.2020 | 84 | −17 | 5 | |
| Andorra | 29.03.2020 | 25.03.2020 | 24.03.2020 | 35 | −4 | −5 | |
| Austria | 19.03.2020 | 13.03.2020 | 22.03.2020 | 85 | −6 | 3 | 7.5 |
| Belgium | 17.03.2020 | 14.03.2020 | 04.04.2020 | 81 | −3 | 18 | 45 |
| Bosnia and Herzegovina | 29.03.2020 | 11.03.2020 | 01.04.2020 | 90 | −18 | 3 | |
| Bulgaria | 28.03.2020 | 13.03.2020 | 29.03.2020 | 73 | −15 | 1 | |
| Croatia | 29.03.2020 | 14.03.2020 | 26.03.2020 | 96 | −15 | −3 | |
| Czechia | 25.03.2020 | 11.03.2020 | 26.03.2020 | 82 | −14 | 1 | |
| Denmark | 19.03.2020 | 11.03.2020 | 31.03.2020 | 72 | −8 | 12 | 3.9 |
| Estonia | 02.04.2020 | 16.03.2020 | 27.03.2020 | 72 | −17 | −6 | 4.6 |
| Finland | 27.03.2020 | 16.03.2020 | 04.04.2020 | 60 | −11 | 8 | 7.7 |
| France | 05.03.2020 | 13.03.2020 | 08.04.2020 | 91 | 8 | 34 | 23 |
| Germany | 13.03.2020 | 16.03.2020 | 26.03.2020 | 73 | 3 | 13 | 5.5 |
| Greece | 19.03.2020 | 12.03.2020 | 26.03.2020 | 84 | −7 | 7 | 2.9 |
| Hungary | 22.03.2020 | 11.03.2020 | 08.04.2020 | 77 | −11 | 17 | 0.2 |
| Iceland | 06.04.2020 | 16.03.2020 | 23.03.2020 | 54 | −21 | −14 | |
| Ireland | 23.03.2020 | 13.03.2020 | 09.04.2020 | 91 | −10 | 17 | |
| Italy | 24.02.2020 | 22.02.2020 | 20.03.2020 | 92 | −2 | 25 | 43 |
| Luxembourg | 21.03.2020 | 13.03.2020 | 22.03.2020 | 80 | −8 | 1 | 17 |
| Netherlands | 13.03.2020 | 12.03.2020 | 05.04.2020 | 80 | −1 | 23 | 34 |
| Norway | 18.03.2020 | 12.03.2020 | 23.03.2020 | 70 | −6 | 5 | 2.6 |
| Poland | 22.03.2020 | 12.03.2020 | 05.04.2020 | 81 | −10 | 14 | 2.8 |
| Portugal | 20.03.2020 | 16.03.2020 | 29.03.2020 | 82 | −4 | 9 | 14 |
| Romania | 23.03.2020 | 09.03.2020 | 09.04.2020 | 87 | −14 | 17 | |
| Serbia | 28.03.2020 | 15.03.2020 | 11.04.2020 | 100 | −13 | 14 | |
| Slovakia | 15.04.2020 | 10.03.2020 | 13.04.2020 | 87 | −36 | −2 | |
| Slovenia | 26.03.2020 | 16.03.2020 | 24.03.2020 | 79 | −10 | −2 | 2.9 |
| Spain | 07.03.2020 | 10.03.2020 | 25.03.2020 | 72 | 3 | 18 | 55 |
| Sweden | 16.03.2020 | 29.03.2020 | 19.04.2020 | 46 | 13 | 34 | 29 |
| Switzerland | 13.03.2020 | 13.03.2020 | 21.03.2020 | 77 | 0 | 8 | 16 |
| United Kingdom | 10.03.2020 | 22.03.2020 | 08.04.2020 | 76 | 12 | 29 | |

*$Str_{idx}$ denotes the Stringency index.*

**Supplementary Figure 9** in the appendix, the method seems to mitigate reporting inaccuracies, by providing an $R_t$ with very similar trend.

## Impact of NPIs

Our analysis shows that when $R_t$, based on the confirmed cases, reduced below one, the median severity of the measures for each category was important with a median stringency index of 79.6 out of 100. In addition, the standard deviation of the index, which is equal to 14.3, shows that most countries required measures with similar intensity achieved through different combinations of NPIs. It is not possible to determine the impact of each individual measure as most countries took them in different order and often a given country took multiples ones at the same time, but the high stringency index reinforces the central idea that only important combinations of NPIs allow to control the pandemic. This finding is consistent with the findings presented in (23) where it is shown that initial NPIs managed to reduce the $R_t$, but that only a full contact ban reduced it below one. It is interesting to analyze the measure individually, not to determine their individual impact, but to determine which set of measures country had put in place when they successfully controlled the epidemic. If we look at the median restrictions when countries managed to control the epidemic, they were all at their maximum level apart from some exceptions on the closing of public transport as well as people being allowed to go out of with minimal daily exceptions. The two categories which had the strongest restrictions were the restrictions on public events and the school closing. All countries required canceling public events apart from Sweden and Andorra which only recommended to cancel them. One limitation of the dataset used in this analysis is that it does not measure whether people have to wear mask either in public

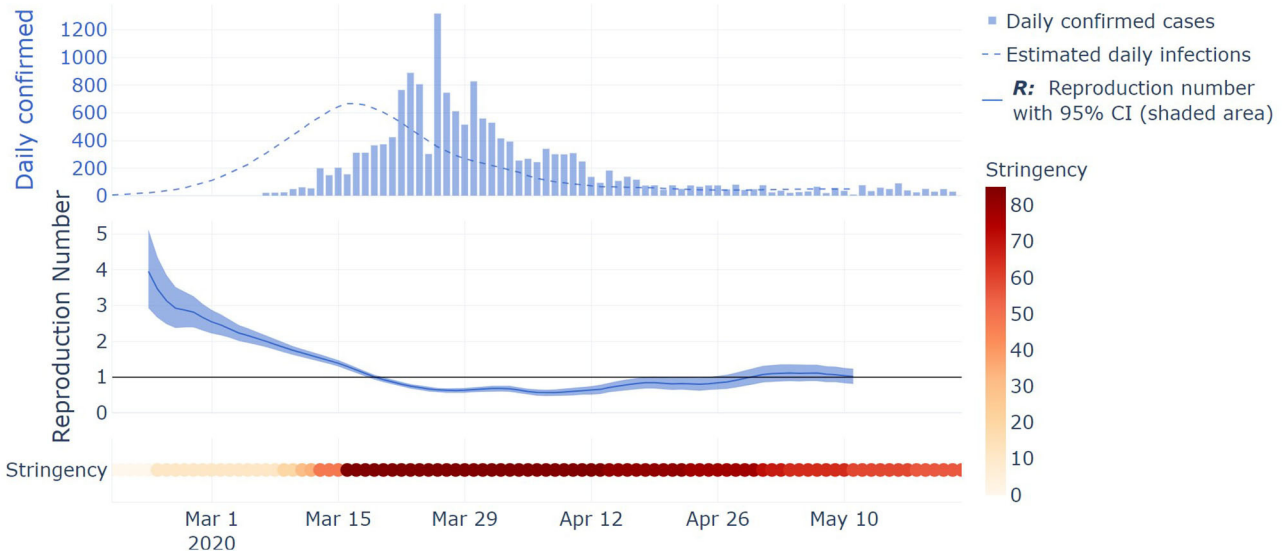



**FIGURE 3 |** Illustration of key steps of the methodology on Austria case study. Top part: histogram of the daily confirmed cases, and inferred daily infections from daily confirmed cases and the derived period distributions displayed as a dashed line. Middle part: the mean estimated $R_t$ is displayed as a full line, along with its 95% CI displayed as a shaded area, with $R_t$ being estimated from the inferred infections. Bottom part: the evolution of the stringency index is displayed with a colorbar changing toward dark red as the stringency score goes toward its maximum value of 100, through the period of interest (from the date of the 5$^{th}$ death up to the 23rd of May).

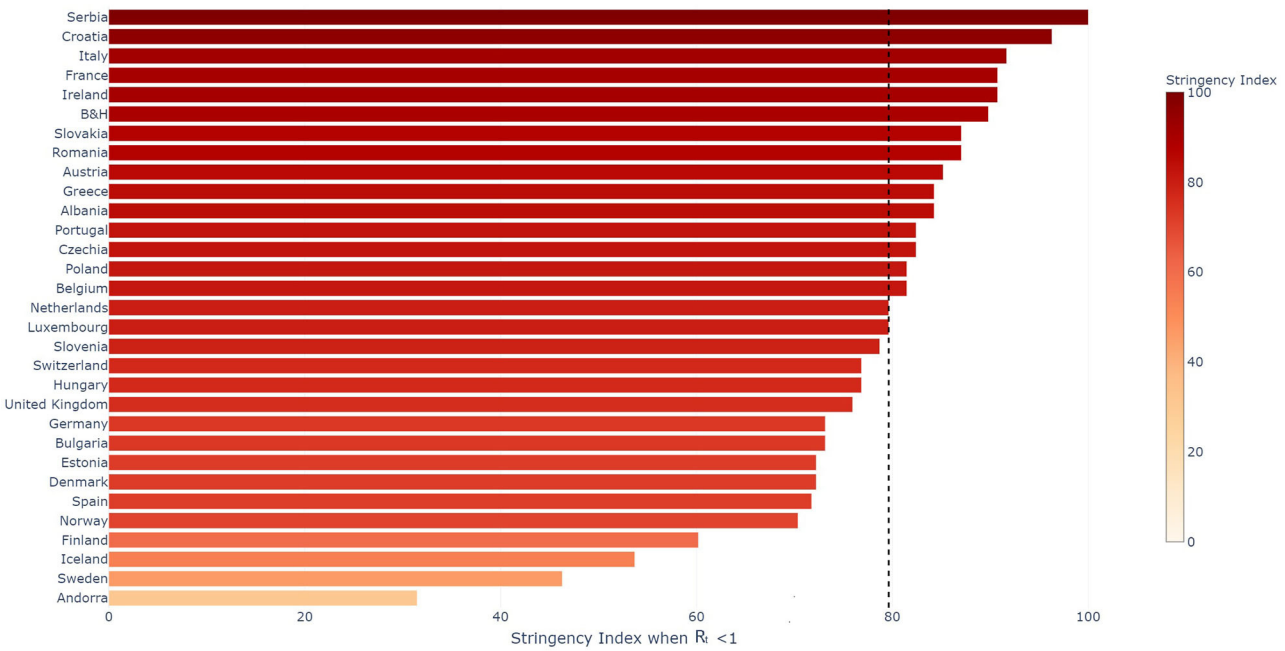



**FIGURE 4 |** Stringency Index per country when $R_t$ evaluated on the confirmed cases reduced below 1. The median value for the set of countries presented in the figure is indicated with the vertical black dotted line.

transport or in all closed environments. It would be important to include those data as more countries are introducing this type of measures to prevent the resurgence of the virus. Also some NPIs could have a higher impact on the mortality, without having a significant impact on $R_t$ evaluated on the confirmed cases. Lastly, the adherence of the population to NPIs is not taken into account here, and is definitely an important parameter to assess their impact on the spread of the pandemic within a country.

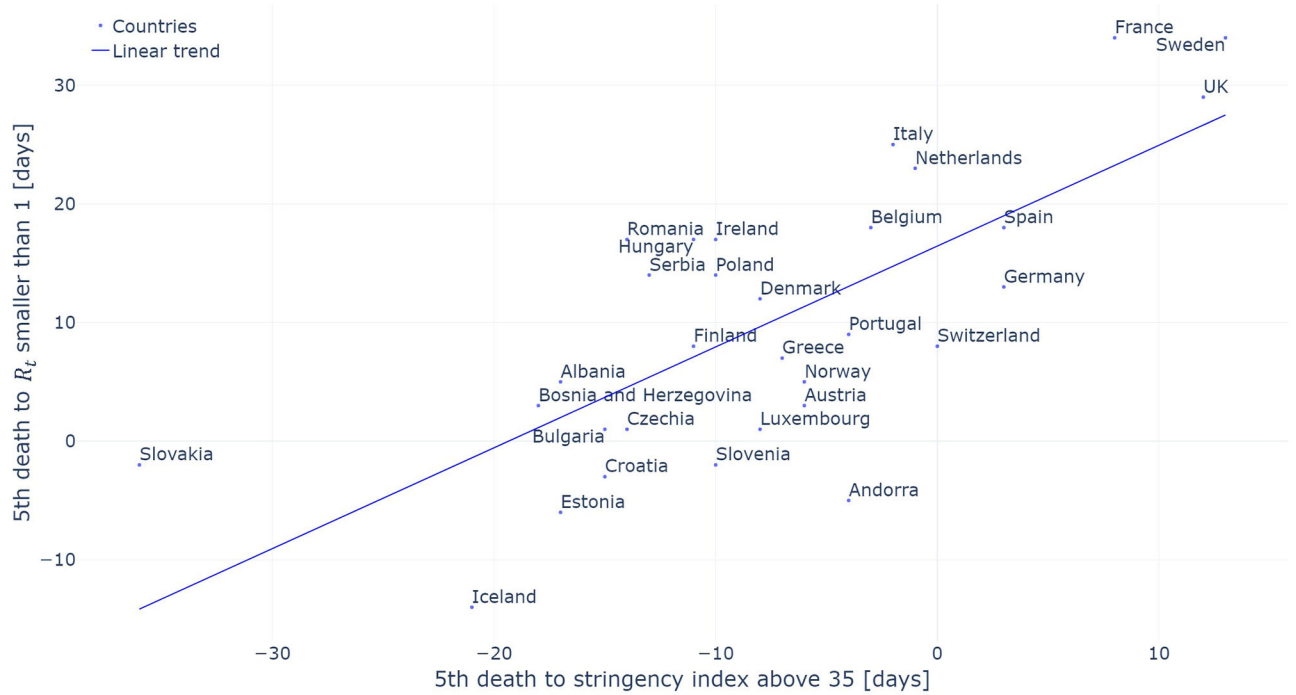



**FIGURE 5 |** Period required to contain the epidemic ($R_t$ smaller than one) evaluated on the confirmed cases as a function of the period between the 5th death and the introduction of initial restrictive measures (stringency index above 35). The linear trend is added for reference.


**FIGURE 6 |** Excess death measured in percent of the number of average death observed between 2015 and 2019 as a function of the period between the 5th death and the introduction of initial restrictive measures (stringency index above 35). The linear trend is added for reference.

Our analysis also looked at the timing of NPIs introduction with the results presented in **Figure 5**. A strong correlation (Pearson coefficient of 0.722) between the time at which NPIs were introduced and the time at which a country managed to reduce $R_t$ below one was found. This correlation indicates that countries which introduced NPIs early on manage to control the evolution of the pandemic within a shorter time frame. The use of the 5th death as a starting date allows to take into account that the pandemic did not start at the same time in all the countries analyzed in this study. The United Kingdom can serve as an interesting example. The UK had initially planned to build "targeted herd immunity" delaying the introduction of restrictive measure. As a result of this delay, the UK was only able to contain the epidemic 29 days after the 5th death occurred in the country when the median time for the countries included in our analysis was of 8 days. There are three outliers in our analysis being Andorra, Sweden and Iceland. Sweden has decided not to introduce a complete lockdown and stands with one of the highest daily death incidence in Europe [May 23rd: Sweden−5.34 deaths per million people per day; other European countries analyzed 0.82 on the same day (34)]. In the preceding analysis, no delay between the application of a measure and its effects on the reproduction number was taken into account. By doing so, the aim is to measure the timing between the introduction of the given measures and its effect on the $R_t$ irrespectively of the behavioral impact it has on the inhabitants who might anticipate the introduction of the measures or inversely take some time to adapt to the introduced measures.

The analysis was replicated using $R_t$ computed on the deaths linked to a SARS-COV2 infections and the data can be found in **Supplementary Figures 11**, **12**. Similarly to the results presented above, countries had a median stringency index of 81.48 when they managed to reduce the $R_t$ computed on deaths below one. It is interesting however to note that the analysis between the introduction of the NPIs and the time at which the $R_t$ reduced below one showed much poorer correlation, Pearsons correlation factor of 0.338, compared to the same analysis on the confirmed cases. A critical limitation when analyzing the evolution of $R_t$ evaluated on reported deaths is that the large variance in the distribution between the onset of the symptoms and the deaths of an individual spreads the retrieved infections. As a results, it becomes very difficult to detect sharp changes in $R_t$ induced by the introduction of NPIs. This effect is similar to the effect of increasing the variance of the incubation period which was shown to decrease the ability to detect changes in $R_t$ (18). The chosen method retrieves the infections dates by subtracting a shift drawn from the distribution of interest. The latter can effectively be seen as a convolution of the confirmed and death cases with the inverse distribution of the corresponding shift, hence spreading the retrieved in time compared to the true level of infections. Using a deconvolution method to retrieve the date of infections instead of the chosen method could improve the detection of changepoints in the trend.

The excess death observed in each country was compared to the timing of the introduction of the NPIs. This analysis has two main benefits. First, it allows to measure the impact of NPIs independently of the estimated $R_t$ and its associated drawbacks described previously. Second, it allows to compare the size of the pandemic in each country without any bias introduced by changes in reporting policy withing a given country which impacts the $R_t$. Such bias would include a rapid increase in the number of tests being performed as tests become more widely available. A Pearson correlation factor of 0.684 between these two variables indicates that countries which took restrictive measures earlier observed lower excess deaths. This high correlation between the timing at which NPIs were introduced and the level of excess death confirms the idea that the $R_t$ evaluated on the confirmed deaths is not appropriate to evaluate the impact of these measures. One bias introduced by using the level of excess deaths to assess the impact of NPIs is that excess deaths will be larger in countries with older populations for a given penetration of the virus in the population as the elderly are much more vulnerable to the virus (45, 46).

A drawback of considering the evolution in the different countries at a national level and not at a regional one is that the heterogeneity of the spread of the virus is disregarded. To evaluate not only the effects of NPIs but also the resurgence of localized clusters, whose identification will be critical to avoid new waves, it is important to look where the cases are located at a more local level. There is therefore a trade-off where $R_t$ is more reliable when evaluated on a larger amount of cases, but less representative as it does not take into account local disparities. Given the greater risk for older population to die or be hospitalized, it would also be interesting to assess the impact of different NPIs across different age groups.

## CONCLUSION

The proposed method to estimate the effective reproduction number $R_t$ has been shown to be less reliant on the initial assumptions of $R_0$ and to effectively improve the modelization of discontinuities in $R_t$ which could be for example observed near the introduction of NPIs. The developed method was subsequently used to analyze the impact of NPIs on 31 European countries. It was first demonstrated that during the first semester of 2020, most European countries had to implement important restrictions to control the pandemic. Our analysis was further extended to show that early introduction of NPIs shortened the time required to control the evolution of the pandemic. The latter correlation was validated by highlighting a direct correlation between early adoption of restrictive measures and a reduction in the excess deaths.

Our study on the impact of health measures focused on European countries but can be extended to other countries for which data on the daily incidence as well as the NPIs taken on a given day are available. To extend this study to a larger set of countries, it would however be necessary to adapt the period between the onset of the symptoms and a case being confirmed or the death of a patient. However, while a sensitivity analysis would be required to assert the influence of variations in the different period distributions, the relatively small difference between the periods derived in Switzerland and in China (6.3%) in regards to the incertitude on the other parameters (daily incidence, infectivity profile) lets us believe that this factor is

likely to play a marginal role if our analysis was to be extended to more countries.

Additional data could help refining our conclusions. First, we could add hospitalizations data as those would not be influenced by change in testing policies within a given country. In addition, looking at $R_t$ within the different age groups could improve our understanding of the impacts of the different NPIs on these various groups. This information would be crucial to develop effective health policies protecting the most vulnerable while provoking minimal disruptions to the society and the economy.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The excess death was retrieved from: Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina, and Joe Hasell, Coronavirus Pandemic (COVID-19), Our World in Data, https://ourworldindata.org/coronavirus [accessed on 24/05/2020]. Data related to the period between a positive test and the death of an individual were retrieved from: Swiss Federal Office of Public Health (FOPH), Cas confirmés en laboratoire: distribution géographique, https://covid-19-schweiz.bagapps.ch/fr-1.html [accessed on 06/05/2020] $R_t$ was estimated using incidence data for confirmed cases and death published in the COVID-19 Data Repository by Johns Hopkins University (JHU CSSE), https://github.com/CSSEGISandData/COVID-19 [accessed on 23/05/2020]. Data regarding the various state interventions were retrieved from the Coronavirus government response tracker (OxCGRT) developed by the Blavatnik School of Government, Oxford University, https://github.com/OxCGRT/covid-policy-tracker [accessed on 23/05/2020].

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

HT performed the analysis of the data as well as the redaction of the article. MB contributed to the analysis as well as the redaction of the article. AR reviewed the method used to analyze the data. CG-B extracted the data used for the analysis. J-PG reviewed the article. CL initiated the project and the research question and contributed to the redaction of the article. All authors contributed to the article and approved the submitted version.

## ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at medRxiv under the title Adaptive time-dependent priors and Bayesian inference to evaluate SARS-CoV-2 public health measures validated on 31 countries (47).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2020.583401/full#supplementary-material

## REFERENCES

1. Mandal M, Jana S, Nandi SK, Khatua A, Adak S, Kar TK. A model based study on the dynamics of COVID-19: prediction and control. *Chaos Solitons Fractals.* (2020) 136:109889. doi: 10.1016/j.chaos.2020.109889
2. Huang L, Zhang X, Zhang X, Wei Z, Zhang L, Xu J, et al. Rapid asymptomatic transmission of COVID-19 during the incubation period demonstrating strong infectivity in a cluster of youngsters aged 16-23 years outside Wuhan and characteristics of young patients with COVID-19: a prospective contact-tracing study. *J Infect.* (2020) 80:e1–3. doi: 10.1016/j.jinf.2020.03.006
3. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the basic reproduction number (R0). *Emerg Infect Dis.* (2019) 25:1–4. doi: 10.3201/eid2501.171901
4. Britton T, Ball F, Trapman P. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science.* (2020) 369:846–9. doi: 10.1126/science.abc6810
5. Thomas LJ, Huang P, Yin F, Luo XI, Almquist ZW, Hipp JR, et al. Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity. *Proc Natl Acad Sci USA.* (2020) 117:24180–87. doi: 10.1073/pnas.2011656117
6. Wang Y, Teunis P. Strongly heterogeneous transmission of COVID-19 in Mainland China: local and regional variation. *Front Med.* (2020) 7:329. doi: 10.3389/fmed.2020.00329
7. van den Driessche P. Reproduction numbers of infectious disease models. *Infect Dis Model.* (2017) 2:288–303. doi: 10.1016/j.idm.2017.06.002
8. Perasso A. An introduction to the basic reproduction number in mathematical epidemiology. *ESAIM: ProcS.* (2018) 62:123–38. doi: 10.1051/proc/201862123
9. Cintrón-Arias A, Castillo-Chávez C, Bettencourt LMA, Lloyd AL, Banks HT. The estimation of the effective reproductive number from disease outbreak data. *Math Biosci Eng.* (2009) 6:261–82. doi: 10.3934/mbe.2009.6.261
10. Bettencourt LMA, Ribeiro RM. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS ONE.* (2008) 3:e2185. doi: 10.1371/journal.pone.0002185
11. Riley S. Transmission dynamics of the etiological agent of sars in Hong Kong: impact of public health interventions. *Science.* (2003) 300:1961–6. doi: 10.1126/science.1086478
12. Gani R, Leach S. Transmission potential of smallpox in contemporary populations. *Nature.* (2001) 414:748–51. doi: 10.1038/414748a
13. King AA, Domenech de Cellès M, Magpantay FMG, Rohani P. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc R Soc B.* (2015) 282:20150347. doi: 10.1098/rspb.2015.0347
14. Byambasuren O, Cardona M, Bell K, Clark J, McLaws ML, Glasziou P. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. *Off J Assoc Med Microbiol Infect Dis Can.* (2020). doi: 10.3138/jammi-2020-0030. [Epub ahead of print].
15. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol.* (2004) 160:509–16. doi: 10.1093/aje/kwh255
16. Hens N, Van Ranst M, Aerts M, Robesyn E, Van Damme P, Beutels P. Estimating the effective reproduction number for pandemic influenza from notification data made publicly available in real time: a multi-country analysis for influenza A/H1N1v 2009. *Vaccine.* (2011) 29:896–904. doi: 10.1016/j.vaccine.2010.05.010

17. Cauchemez S, Boëlle PY, Thomas G, Valleron AJ. Estimating in real time the efficacy of measures to control emerging communicable diseases. *Am J Epidemiol.* (2006) 164:591–7. doi: 10.1093/aje/kwj274
18. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am J Epidemiol.* (2013) 178:1505–12. doi: 10.1093/aje/kwt133
19. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health.* (2020) 5:e261–70. doi: 10.1016/S2468-2667(20)30073-6
20. Zhang Y, Jiang B, Yuan J, Tao Y. The impact of social distancing and epicenter lockdown on the COVID-19 epidemic in mainland China: a data-driven SEIQR model study. *medRxiv [Preprint].* (2020). doi: 10.1101/2020.03.04.20031187
21. Davies NG, Kucharski AJA, CMMID Eggo RM, Gimma COVID-19 Working Group, Edmunds WJ. The effect of non-pharmaceutical interventions on COVID-19 cases, deaths and demand for hospital services in the UK: a modelling study. *Lancet Public Health.* (2020) 5:E375–85. doi: 10.1016/S2468-2667(20)30133-X
22. Khailaie S, Mitra T, Bandyopadhyay A, Schips M, Mascheroni P, Vanella P, et al. Estimate of the development of the epidemic reproduction number Rt from Coronavirus SARS-CoV-2 case data and implications for political measures based on prognostics. *medRxiv [Preprint].* (2020). doi: 10.1101/2020.04.04.20053637
23. Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, Wilczek M, et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science.* (2020) 369:eabb9789. doi: 10.1126/science.abb9789
24. Roux J, Massonnaud C, Crépey P. COVID-19: one-month impact of the French lockdown on the epidemic burden. *medRxiv [Preprint].* (2020). doi: 10.1101/2020.04.22.20075705
25. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature.* (2020) 584:257–61. doi: 10.1038/s41586-020-2405-7
26. Brzezinski A, Deiana G, Kecht V, Dijcke D. *The COVID-19 Pandemic: Government vs. Community Action Across the United States*. INET Oxford Working Paper No: 2020–06 (2020).
27. Abbott S, Hellewell J, Thompson RN, Sherratt K, Gibbs HP, Bosse NI, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res.* (2020) 5:112. doi: 10.12688/wellcomeopenres.16006.1
28. Gostic KM, McGough L, Baskerville E, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, Rt. *medRxiv [Preprint].* (2020). doi: 10.1101/2020.06.18.20134858
29. Marschner IC. Back-projection of COVID-19 diagnosis counts to assess infection incidence and control measures: analysis of Australian data. *Epidemiol Infect.* (2020) 148:e97. doi: 10.1017/S0950268820001065
30. Mieskolainen M, Bainbridge R, Buchmueller O, Lyons L, Wardle N. Statistical techniques to estimate the SARS-CoV-2 infection fatality rate. *medRxiv [Preprint].* (2020). doi: 10.1101/2020.11.19.20235036
31. Becker NG, Watson LF, Carlin JB. A method of non-parametric back-projection and its application to aids data. *Statist Med.* (1991) 10:1527–42. doi: 10.1002/sim.4780101005
32. Petermann M, Wyler D. A pitfall in estimating the effective reproductive number Rt for COVID-19. *Swiss Med Wkly.* (2020) 150:w20307. doi: 10.4414/smw.2020.20307
33. Johns Hopkins University. *COVID-19 Data Repository*. Johns Hopkins University. (2020). Available online at: https://github.com/CSSEGISandData/COVID-19 (accessed May 23, 2020).
34. Roser M, Ritchie H, Ortiz-Ospina E, Hasell J. *Coronavirus Pandemic (COVID-19) - Our World in Data.* (2020). Available online at: https://github.com/owid/covid-19-data (accessed May 26, 2020).
35. Swiss Federal Office of Public Health. *Cas Confirmés en Laboratoire: Distribution Géographique*. FOPH. (2020). Available online at: https://covid-19-schweiz.bagapps.ch/fr-1.html (accessed June 5, 2020).
36. Hale T, Webster S, Petherick A, Philips T, Kira B. *Coronavirus Government Response Tracker (OxCGRT)*. (2020). Available online at: https://github.com/OxCGRT/covid-policy-tracker (accessed May 23, 2020).
37. Hale T, Petherick, A, Philips T, Webber S. Variation in government responses to COVID-19. *Blavatnik School of Government Working Paper*. Version 5.0. (2020). Available online at: www.bsg.ox.ac.uk/covidtracker
38. Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. *Int J Infect Dis.* (2020) 93:284–86. doi: 10.1016/j.ijid.2020.02.060
39. Linton NM, Kobayashi T, Yang Y, Hayashi K, Jung S, Yuan B, et al. Epidemiological characteristics of novel coronavirus infection: a statistical analysis of publicly available case data. *medRxiv [Preprint].* (2020). doi: 10.1101/2020.01.26.20018754
40. Scire J, Nadeau S, Vaughan T, Brupbacher G, Fuchs S, Sommer J, et al. Reproductive number of the COVID-19 epidemic in Switzerland with a focus on the Cantons of Basel-Stadt and Basel-landschaft. *Swiss Med Wkly.* (2020) 150:w20271. doi: 10.4414/smw.2020.20271
41. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis.* (2020) 20:669–77. doi: 10.1016/S1473-3099(20)30243-7
42. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J Travel Med.* (2020) 27:taaa021. doi: 10.1093/jtm/taaa021
43. Bastos LS, Economou T, Gomes MFC, Villela DAM, Coelho FC, Cruz OG, et al. A modelling approach for correcting reporting delays in disease surveillance data. *Stat Med.* (2019) 38:4363–77. doi: 10.1002/sim.8303
44. Günther F, Bender A, Katz K, Küchenhoff H, Höhle M. Nowcasting the COVID-19 pandemic in Bavaria. *Biometric J.* (2020). doi: 10.1002/bimj.202000112. [Epub ahead of print].
45. Ruan S. Likelihood of survival of coronavirus disease 2019. *Lancet Infect Dis.* (2020) 20:630–1. doi: 10.1016/S1473-3099(20)30257-7
46. Perez-Saez J, Lauer SA, Kaiser L, Regard S, Delaporte E, Guessous I, et al. Serology-informed estimates of SARS-CoV-2 infection fatality risk in Geneva, Switzerland. *Lancet Infect Dis*. (2020). doi: 10.1016/S1473-3099(20)30584-3
47. Turbé H, Bjelogrlic M, Robert A, Gaudet-Blavignac C, Goldman J-P, Lovis C. Adaptive time-dependent priors and Bayesian inference to evaluate SARS-CoV-2 public health measures validated on 31 countries. *medRxiv [Preprint].* (2020). doi: 10.1101/2020.06.10.20126870


**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2021 Turbé, Bjelogrlic, Robert, Gaudet-Blavignac, Goldman and Lovis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Cellular and Infection Microbiology

BRIEF RESEARCH REPORT
published: 15 February 2021
doi: 10.3389/fcimb.2021.610348

# Learning From Limited Data: Towards Best Practice Techniques for Antimicrobial Resistance Prediction From Whole Genome Sequencing Data

*Lukas Lüftinger*[1,2], *Peter Májek*[1], *Stephan Beisken*[1], *Thomas Rattei*[2] *and Andreas E. Posch*[1*]

[1] Ares Genetics GmbH, Vienna, Austria, [2] Division of Computational Systems Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria

OPEN ACCESS

**Edited by:**
Adrian Egli,
University Hospital of Basel, Switzerland

**Reviewed by:**
Helena M. B. Seth-Smith,
University Hospital of Basel, Switzerland
Xiaowei Zhan,
University of Texas Southwestern Medical Center, United States
Samuel A. Shelburne,
University of Texas MD Anderson Cancer Center, United States

***Correspondence:**
Andreas E. Posch
andreas.posch@ares-genetics.com

**Specialty section:**
This article was submitted to Clinical Microbiology, a section of the journal Frontiers in Cellular and Infection Microbiology

**Received:** 25 September 2020
**Accepted:** 11 January 2021
**Published:** 15 February 2021

**Citation:**
Lüftinger L, Májek P, Beisken S, Rattei T and Posch AE (2021) Learning From Limited Data: Towards Best Practice Techniques for Antimicrobial Resistance Prediction From Whole Genome Sequencing Data. Front. Cell. Infect. Microbiol. 11:610348. doi: 10.3389/fcimb.2021.610348

Antimicrobial resistance prediction from whole genome sequencing data (WGS) is an emerging application of machine learning, promising to improve antimicrobial resistance surveillance and outbreak monitoring. Despite significant reductions in sequencing cost, the availability and sampling diversity of WGS data with matched antimicrobial susceptibility testing (AST) profiles required for training of WGS-AST prediction models remains limited. Best practice machine learning techniques are required to ensure trained models generalize to independent data for optimal predictive performance. Limited data restricts the choice of machine learning training and evaluation methods and can result in overestimation of model performance. We demonstrate that the widely used random k-fold cross-validation method is ill-suited for application to small bacterial genomics datasets and offer an alternative cross-validation method based on genomic distance. We benchmarked three machine learning architectures previously applied to the WGS-AST problem on a set of 8,704 genome assemblies from five clinically relevant pathogens across 77 species-compound combinations collated from public databases. We show that individual models can be effectively ensembled to improve model performance. By combining models *via* stacked generalization with cross-validation, a model ensembling technique suitable for small datasets, we improved average sensitivity and specificity of individual models by 1.77% and 3.20%, respectively. Furthermore, stacked models exhibited improved robustness and were thus less prone to outlier performance drops than individual component models. In this study, we highlight best practice techniques for antimicrobial resistance prediction from WGS data and introduce the combination of genome distance aware cross-validation and stacked generalization for robust and accurate WGS-AST.

**Keywords: machine learning, genomics, antimicrobial resistance, antibiotics, whole genome sequencing (WGS)**

## INTRODUCTION

Antimicrobial resistance (AMR) is a rising global threat to human health. To ensure the continued efficacy of antimicrobial compounds, prudent use of this resource is crucial (O'Neill, 2016). Accurate determination of antimicrobial resistance *via* antimicrobial susceptibility testing (AST) is crucial to ensure optimal patient treatment as well as to inform antibiotic stewardship and outbreak monitoring.

In this context, resistance predictions from WGS data may effectively complement phenotypic AST: The time-to-result (TTR) of WGS-based workflows is effectively governed by the continuously decreasing cost and runtime of genome sequencing, while phenotypic testing is ultimately limited by the pathogen's growth rate (Bradley et al., 2015; Břinda et al., 2018). Machine learning (ML) algorithms are increasingly applied for prediction of AMR from WGS data (WGS-AST). Recently described WGS-AST techniques use nucleotide k-mer representations of genome assemblies or raw sequencing data, attempting to learn differences in k-mer counts or presence/absence patterns that correlate with shifts in susceptibility to a target antibiotic (Drouin et al., 2016; Aun et al., 2018; Nguyen et al., 2018a; Drouin et al., 2019). This data-driven approach does not require expert knowledge of AMR mechanisms or prior information on AMR genes, and can thus also be applied towards learning of models for novel antibiotics and unknown resistance mechanisms. Other representations of genomic data, such as amino acid k-mers or protein variants have been used for WGS-AST model training as well (Kim et al., 2020; Valizadehaslani et al., 2020).

Challenges arise, however, when learning is not based on features derived from validated, curated AMR markers for the resistance phenotype in question. For example, the significant impact of population structure when applying ML algorithms to WGS-AST data has been noted before (Hicks et al., 2019). Performance of ML models evaluated on isolates from the same experiment as the training data tends to be significantly higher than performance on isolates sampled from independent data sources. Due to limited availability of WGS data coupled with AST information, the performance of WGS-AST models is usually evaluated by cross-validation (CV). Most commonly this is performed using a random splitting criterion, i.e., by dividing samples randomly (Davis et al., 2016; Nguyen et al., 2018a; Drouin et al., 2019). Performance measures obtained by random CV can however only be assumed valid for the larger population if the sample-generating process yields approximately independent and identically distributed (i.i.d.) samples (Ruppert, 2004). This assumption is violated in data points generated by evolutionary processes, which are correlated as a function of the recency of their last common ancestor. This includes, for example, data pertaining to gene function (Tabe-Bordbar et al., 2018) or protein structure (AlQuraishi, 2019), but also whole genomes. By random splitting, similar samples in an existing dependence structure, e.g., evolutionary distance, may be split into the training and test set of CV. This causes the model to overfit by learning features that are spuriously correlated with the phenotype, features which are also present in the test set due to the violated assumption of independence. (Roberts et al., 2017) For example, k-mers mapping to the replication machinery of a resistance cassette-carrying plasmid vector may be highly correlated with resistance due to the prevalence of the plasmid in resistant isolates, despite not contributing to resistance itself. A model overfit to this population by inclusion of such spurious correlations may fail unexpectedly on a population of isolates where the resistance cassette has integrated into the genome. Biological datasets with low sample count but a high number of features further increase the potential of dependence structures and the risk of overfitting (Clarke et al., 2008), and are known to be susceptible to overestimation of model performance by random CV (Roberts et al., 2017).

Ultimately, applying a trained model to multiple large and independently sampled datasets is the gold standard for gauging model generalizability, though this is currently impractical for WGS-AST. To estimate generalization performance in the absence of additional data, blocking CV techniques can be used. Blocking CV seeks to split data into pre-defined similar groups of samples, thus reducing the splitting of dependence structures into the training and test sets of CV (Valavi et al., 2019).

Another significant challenge in achieving robust WGS-AST models with high predictive accuracy is selection of an appropriate learning algorithm. High dimensionality and a low number of training samples constrain the selection of suitable choices. In this study we selected three established learning algorithms which have previously been applied to the WGS-AST problem, and exhaustively benchmarked them across a set of five clinically relevant pathogens (*A. baumannii, E. coli, K. pneumoniae, P. aeruginosa and S. aureus*) and a total of 77 species-compound combinations. We also investigated the possibility of improving model accuracy and robustness by ensembling different learning algorithms such as majority vote and stacked generalization (Wolpert, 1992). This commonly used set of techniques has, to the best of our knowledge, not been explored in the context of antimicrobial resistance prediction from WGS data.

## RESULTS

### Random CV May Overestimate WGS-AST Model Generalizability

To assess the impact of data splitting techniques on performance estimates of WGS-AST models, we trained extreme gradient boosting (Chen and Guestrin, 2016) models under random and genome distance-aware CV. Genome distance-aware CV attempts to improve independence of test sets by segregating samples based on a known dependence structure in the data, namely genome similarity (see Methods). This mirrors the application of the trained model towards independently sampled datasets, in the absence of actual new data.

Genome assemblies coupled with AST information were obtained from public databases (see Methods) for five human

pathogens (*A. baumannii*, *E. coli*, *P. aeruginosa*, *K. pneumoniae* and *S. aureus*) and a total set of 77 organism/compound combinations. Data was split into 5 CV folds by either a random or genome distance-aware splitting criterion. Random CV splitting was repeated 10 times while varying the random seed to enable significance estimation (see **Supplementary Methods Section 3**). Extreme gradient boosting (XGB) machine learning models were trained on nucleotide k-mer representations of each of the resulting training sets (see Methods) and evaluated on the corresponding test sets.

Of the 77 investigated organism/compound pairs, 60 exhibited significantly higher balanced accuracy (bACC) estimates for random CV than for genome distance-aware CV (**Figure 1**). The average bACC estimated by random CV was 4.45% greater than that of distance-aware CV, indicating that performance estimates by random CV are likely to overestimate the true performance of WGS-AST models on unseen, independent data sampled from a population that is not comprehensively represented in the training data. The observed effect is congruent with published findings of the generalization properties of WGS-AST models applied to independently sampled data (Hicks et al., 2019). To empirically demonstrate that performance estimates by random CV are prone to be overoptimistic we trained XGB models on the full set of *P. aeruginosa* samples and evaluated them on an independent dataset of 140 samples (Ferreira et al., 2020) (see **Supplementary Figure S1**). On average, bACC of the trained XGB models on this test set was 10.12% lower than estimated by random CV. Distance-aware CV provided more conservative estimates while not completely rescuing the overestimation bias, likely due to novel AMR mechanisms associated with the independent dataset (see *Discussion*).

## Benchmarking of Machine Learning Algorithms for WGS-AST

We selected three machine learning algorithms for prediction of antimicrobial resistance from WGS data represented as nucleotide k-mer profiles: extreme gradient boosting (XGB) (Chen and Guestrin, 2016), elastic net regularized logistic regression (ENLR) (Friedman et al., 2010), and set covering machine (SCM) (Marchand and Shawe-taylor, 2000). All selected algorithms were recently reported to perform well on the WGS-AST task (Aun et al., 2018; Nguyen et al., 2018a; Drouin et al., 2019; Ferreira et al., 2020; Lees et al., 2020).

Selected algorithms were benchmarked across a set of five clinically relevant bacterial pathogens and a total of 77 organism/compound combinations (**Figure 2A**). Predictive performance across evaluated algorithms was similar, with a median difference between the strongest and weakest model for an organism/compound combination of 4.22% bACC (**Figure 2B**). ENLR, XGB, and SCM algorithms yielded the model with the highest bACC for 34, 28, and 15 datasets, respectively. Despite their characteristically low complexity and high interpretability, SCM models outperformed the more complex ENLR and XGB models on several datasets, particularly when few resistant isolates were available (**Figure 2C**).

## Model Stacking Improves Predictive Performance and Robustness of Individual ML Models

To improve predictive performance, we then employed stacking, a model ensembling technique. The ENLR algorithm was used to train a metamodel which learned to optimally combine predictions of individual component XGB, ENLR and SCM

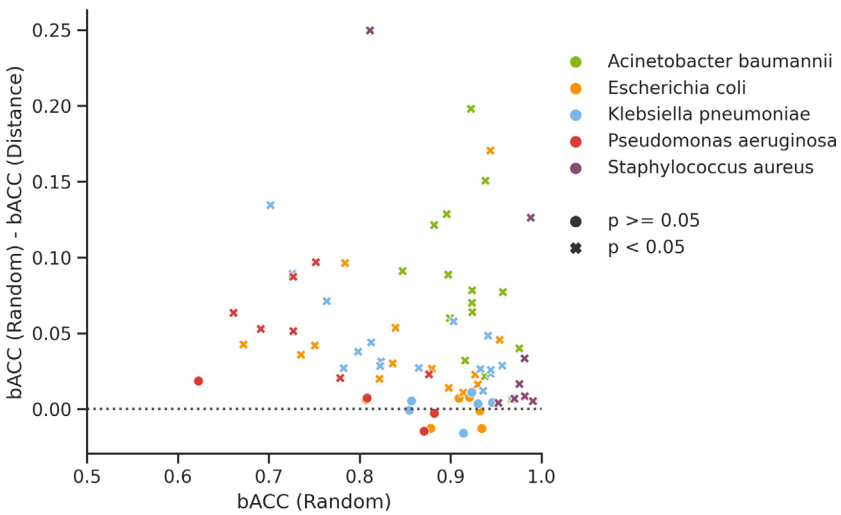



**FIGURE 1** | Difference in balanced accuracy (bACC) of XGB models trained and evaluated under random CV and genome distance-aware CV for all considered organism/compound pairs. Significance thresholds are the probability of obtaining bACC estimates as low or lower than the ones from genome distance-aware CV when sampling from a normal distribution fitted to 10 random CV replicates obtained with different random seeds.

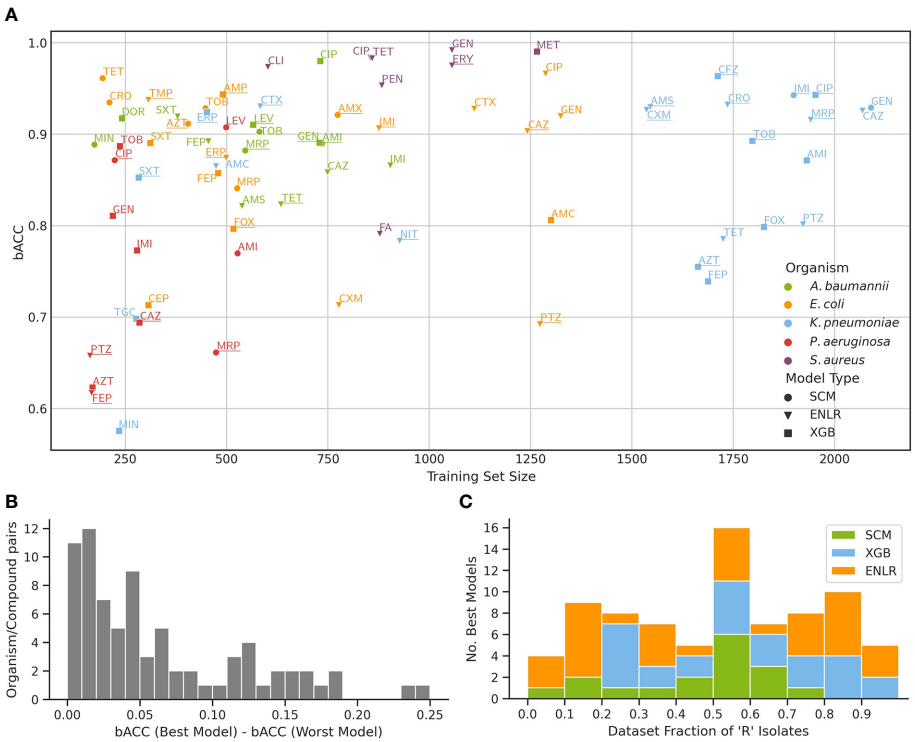



**FIGURE 2** | Benchmark of three ML algorithms on the prediction of antimicrobial resistance from WGS data. **(A)** Predictive performance of models for each organism/compound pair as a function of training set size. For each pair, performance of a model with the highest bACC is shown, and underlined if the stacking model outperformed it. The mapping of compound names to compound abbreviations is given in **Supplementary Table S4**. **(B)** Distribution of bACC differences between the models with highest and lowest bACC for all organism/compound pairs. **(C)** Number of top performing models from each algorithm as a function of the fraction of resistant isolates in the training set.

models (**Figure 3** and Methods). We compared the stacked model with a simpler ensembling approach based on the majority vote of all component models. On average, stacked models improved over the sensitivity and specificity of their component models by 1.77% and 3.20%, respectively. The stacking model was found to be the best model by bACC of outer CV in 30 out of 77 organism/compound combinations, outperforming individual component models and the majority vote ensemble. To gauge robustness, we considered a model to have encountered a failure mode if it exhibited a drop in bACC of more than 5.00% compared to the best model for that organism and compound. The stacked models encountered failure modes in 3 out of 77 cases, thus exhibiting superior robustness compared to component models and the simple majority vote ensemble (**Table 1**).

## Failure Modes of Component Models and Biological Interpretation

We selected two organism/compound pairs with large differential performance among component models and investigated the biological underpinnings of observed failure modes by annotating k-mers mapping to known AMR biomarkers (Ferreira et al., 2020). For practical reasons, we investigated the models trained in the CV fold exhibiting the largest differential performance and considered only the top 10 most impactful features of each model (see **Supplementary Tables 8** and **9**).

For the combination agent piperacillin and tazobactam (PTZ) in *Klebsiella pneumoniae*, the SCM model exhibited a drop of on average 10% bACC in comparison to XGB and ENLR models. This drop was due to decreased specificity of the SCM model, caused by the model making a comparably larger number of false resistance calls (see **Supplementary Table 6**). Of the four features learned by the model, two mapped to known AMR markers *gyrA* and *catB3*, involved in fluoroquinolone and phenicol resistance, respectively, with no known function in PTZ resistance (Bunny et al., 1995; Drlica and Zhao, 1997). This indicates a strong reliance of the model on features which are spuriously correlated with the phenotype. Conversely, the corresponding XGB model learned multiple k-mers mapping to *blaKPC* beta-lactamase genes, known to confer resistance to piperacillin (Bush and Jacoby, 2010). The stacking model incorporating this SCM model learned to fully disregard the predictions of the SCM model in favor of ENLR and XGB predictions (see **Supplementary Table 7**).

Conversely, for tobramycin (TOB) in *Acinetobacter baumannii*, XGB and ENLR exhibited reduced bACC, mostly due to failure to identify resistant samples in one CV fold. The

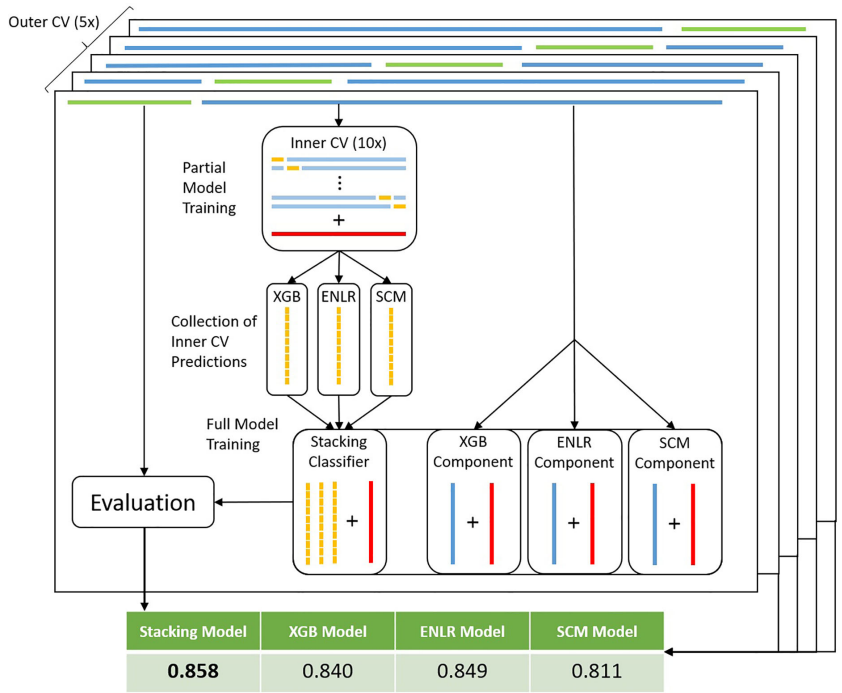



**FIGURE 3** | Workflow for model stacking with nested CV. For each training data set in the outer CV loop (dark blue bars on top) complete with true resistance status of samples (red bars), an inner CV loop is run (light blue bars). The full set of predictions (yellow bars) obtained from the test sets of the inner CV are used to train a stacking model to ideally combine predictions from each of the components. At the same time, full component models are trained on the training data set (blue bars within component models). Subsequently, predictions are made by all full component models on the test dataset (green bars on top). Predictions are made by the stacking model using the component model predictions as input features. Finally, performance metrics are obtained by scoring predictions of each model type against the true resistance status of test set samples.

**TABLE 1** | Summary statistics of model performance (averaged over organisms and compounds) and number of top-1 placements and failure modes (a more than 5% drop in bACC compared to the best performing model) per organism and compound combination.

| Algorithm | bACC | Sensitivity | Specificity | # Top-1 Rankings (bACC) | # Failure Modes Encountered |
|---|---|---|---|---|---|
| *ENLR* | 0.849 | 0.807 | **0.890** | 15 | 9 |
| *XGB* | 0.840 | 0.813 | 0.866 | 13 | 13 |
| *SCM* | 0.811 | 0.805 | 0.818 | 8 | 31 |
| *Majority vote* | 0.846 | 0.818 | 0.873 | 17 | 10 |
| *Stacking* | **0.858** | **0.826** | **0.890** | **30** | **3** |

*Best metrics in boldface.*

SCM model performed consistently well. Feature analysis showed that two of the only three features considered by the SCM model could be mapped to N-Acetyltransferase genes *aadB* and *aacC5*, known to confer resistance to aminoglycosides (Shaw et al., 1993; Cox et al., 2015). The XGB and ENLR models learned a high number of features (512 and 6351, respectively), indicating potential overfitting. In the top 10 features of each, only XGB exhibited interpretable features, namely *aacA16*, an aminoglycoside acetyltransferase, and *msrE*, conferring resistance to erythromycin (Sharkey and O'Neill, 2018). The stacking model learned to assign the highest weight to the SCM component, thereby achieving second place performance after the individual SCM itself (see **Supplementary Table 7**).

# DISCUSSION

## Random CV May Overestimate WGS-AST Model Generalizability

We demonstrate on a large collection of public datasets that special care must be taken when applying machine learning techniques to the WGS-AST problem. Two common properties of genomics datasets, namely high dimensionality (Clarke et al., 2008) and sparse and biased sampling of the underlying data distribution, invalidate default design choices such as random dataset partitioning for evaluation of generalizability.

Awareness of the issue of splitting data for WGS-AST ML is developing; a recent study (Aytan-Aktug et al., 2020) used

genome clustering based on a similarity threshold, splitting only full clusters into different CV folds together. This approach to data partitioning is also widely used in gene- and protein-based deep learning, where generally only a single training, validation, and test dataset are used (AlQuraishi, 2019; Strodthoff et al., 2019). While grouping by a similarity threshold increases biological meaningfulness and independence of data splits (potentially further reducing performance overestimation), it may cause strongly disbalanced CV fold sizes, especially in a small data regime. The genomic distance-aware method proposed in this work by design generates equally sized folds and aims at maximizing the sample independence across the folds. **Supplementary Figure S3** shows how the proposed method partitions public *P. aeruginosa* samples used in this work.

Similarly, hierarchical clustering has been used for removal of highly clonal genomes from the dataset (Nguyen et al., 2018b), though mainly due to computational considerations. While deduplication is likely to reduce the impact of dependence structures in the training data, the large dimensionality and sparsity of AMR information in a genome represented as k-mer counts makes finding a useful deduplication criterion tricky, especially if the goal is for the model to learn unknown AMR mechanisms.

Of note, data splitting methods controlling for population structure are expected to provide performance estimates differing from random splitting under two conditions: significant population structure must exist in the training dataset, and causal AMR mechanisms must be correlated with population structure. Datasets of closely related samples (not reflecting the true diversity of the underlying population), and datasets containing homogeneously distributed AMR mechanisms, allow only limited insight into possible performance drops due to novel AMR mechanisms associated with distinct populations. Thus, such techniques may still overestimate performance on independently sampled datasets to varying degrees.

Ultimately, a comprehensive assessment of the impact of different clustering and deduplication strategies on model generalizability estimates may be valuable. However, to not only overcome overestimation of performance but to raise predictive accuracy beyond FDA requirements for AST devices (FDA, 2009) and hasten application of WGS-AST models in a diagnostic setting, a greater depth and width of training and test data will be required.

## Benchmarking of Machine Learning Algorithms for WGS-AST

Comparing three different ML algorithms, we find that no single algorithm is clearly superior using the respectively chosen feature space, model parametrization and evaluation criteria. While training set size was positively correlated with performance of all investigated algorithms (see **Supplementary Figure S2**), both species identity and antibiotic compound class clearly influenced classifier performance. Previously established findings regarding the significant challenge in providing accurate AMR predictions for *P. aeruginosa* have been affirmed by this work (Aun et al., 2018). Likewise, we obtain high accuracy predictions for *S. aureus* and most antibiotic compounds in *E. coli*, reflecting earlier results obtained with approaches operating on curated sets of AMR markers instead of nucleotide k-mers (Bradley et al., 2015; Moradigaravand et al., 2018). A notable example of the influence of the compound class on prediction accuracy is the consistently high performance of models for resistance to the fluoroquinolones ciprofloxacin (CIP) and levofloxacin (LEV), which is strongly determined by single nucleotide polymorphisms to the DNA gyrase gene *gyrA* and topoisomerase IV gene *parC* (Jacoby, 2005).

## Model Stacking Improves Predictive Performance and Robustness of Individual ML Algorithms

Several WGS-AST machine learning techniques have been described in the scientific literature. We demonstrate that individual ML algorithms, while performing similarly on average, are susceptible to different failure modes when applied to the WGS-AST problem, such that no single algorithm is clearly preferable for all organism and compound combinations. We illustrate that a stacking ensemble improves predictive performance and robustness, largely beyond that of any of its component models.

It has been suggested that the use of a diverse set of learning algorithms improves predictive accuracy of ensembling models (Kuncheva and Whitaker, 2003). While we systematically benchmarked three algorithms previously reported to perform well on the problem at hand, adding additional ML architectures to the stack is straightforward and may be a promising next step to further improve predictive accuracy and robustness, even in the absence of additional data. Conversely, we note that in settings where model interpretability is of overriding importance, for example in biomarker discovery, individual highly interpretable models such as the SCM may be preferred over complex model ensembles.

## Conclusion

We describe the choice of ML model evaluation strategy and architecture as key aspects affecting model performance and generalizability based on publicly available WGS-AST data sets. To facilitate WGS-AST across organism-compound combinations and translation into clinical practice, applying best practice machine learning techniques and further complementing of publicly available WGS-AST data is important.

# MATERIALS AND METHODS

## Data Retrieval

Genome assemblies and associated resistance/susceptibility profiles for five clinically relevant pathogens (*A. baumannii, E. coli, K. pneumoniae, P. aeruginosa*, and *S. aureus*) were obtained from public data sources (See **Supplementary Tables 1** and **2**) (Karp et al., ; NCBI NCBI, ; Kos et al., 2015; Wattam et al., 2016; Nguyen et al., 2018a; Mahfouz et al., 2020). Minimum inhibitory

concentration (MIC) values, if present, were interpreted (S/I/R) *via* clinical breakpoints according to CLSI 29 standards (Wayne, 2019). Intermediate phenotypes were treated as resistant for model training and evaluation. Isolates with MIC values less than or equal to a dilution step in the intermediate range (meaning that the MIC interpretive category was ambiguous according to CLSI 29 standards) were treated as susceptible. Data was filtered to pass assembly QC metrics (Ferreira et al., 2020). Finally, only organism-compound pairs were included for which at least 50 susceptible and resistant isolates as well as 200 isolates in total could be retrieved (see **Supplementary Tables 1–3**). Using these cut-offs, a total number of 8704 genome assemblies were retrieved.

Genome assemblies used for evaluation of CV estimates on an independent dataset (Ferreira et al., 2020) were obtained from NCBI (PRJNA553678). AST data were obtained from the authors.

## Data Partitioning for Training and Evaluation

Models were trained and evaluated in a nested 10x/5x cross-validation scheme, whereby the inner 10x cross-validation was used to obtain the training features for the stacking model (**Figure 2**).

Genome-distance-based cross-validation folds were created for each species individually such that genome distance was maximized between the test sets of folds (see **Supplementary Methods Section 1**). In short, for all assemblies of each organism, a distance matrix was computed with Mash v2.2 (Ondov et al., 2016). From the distance matrix, two seed samples with the largest genomic distance among them were identified. Subsequently, for each remaining sample, the minimal distance to either of the seeds was computed. Additional seed samples up to the number of desired CV folds were added by selecting samples with the highest minimal distance to existing seeds. Finally, all remaining samples were assigned to seed samples iteratively by assigning to each seed the sample with the lowest genomic distance. The generated five sample groups of even size were used as input to CV. Randomly split CV folds for comparison were created using scikit-learn (Pedregosa et al., 2011).

## Feature Creation and Feature Selection

For XGB and ENLR models, feature extraction and selection were performed according to the following procedure. For all training assemblies of each organism, a count matrix of overlapping k-mers of length 15 was built using KMC 3.1.0 (Kokot et al., 2017). Zero-variance k-mers were removed. Out of all k-mers having identical count profiles across training isolates, only a single representative k-mer was retained. Subsequently, for each organism and relevant antimicrobial compound, a subset of the organism's full count matrix for which S/R class information of the given compound was available was extracted. The k-mer feature space was then condensed by univariate feature selection before application of machine learning. K-mers were tested for independence from the S/R category using the $\chi^2$ test as implemented in scikit-learn and filtered by a p-value of $p < 0.05$. Of the k-mers passing this filtering step, at most 1.5 million k-mers with the highest log-odds ratio were retained. For SCM models, k-mer features of length 31 were created from assemblies with Kover2 according to the supplied manual. To exclude the possibility of biases introduced by common feature selection on the full dataset, features for prediction on the test sets of the outer cross-validation were created only at prediction time.

## Model Training

We trained extreme gradient boosting (XGB), elastic net regularized logistic regression (ENLR) and set covering machine (SCM) models for prediction of antimicrobial susceptibility from WGS data for a set of five clinically relevant pathogens. A fixed set of hyperparameters was used across all organisms and compound pairs, except for the number of trees in the model which was tuned *via* internal CV. We explored the choice of CV method for hyperparameter optimization and found that the performance estimated by the outer CV method is relatively insensitive to the choice of the inner CV method (see **Supplementary Figure S4**) and thus used a distance-based splitting criterion for internal CV of both XGB and ENLR methods. ENLR models were trained using the glmnet_python package, version 0.2.0 (Friedman et al., 2010), and the hyperparameters lambda and alpha were tuned *via* an internal CV. Set covering machine models were trained with the Kover2 package, version 2.0.3 (Drouin et al., 2019) according to the supplied manual and using risk-bound hyperparameter selection (see **Supplementary Methods Sections 4** and **5**).

Individual models were combined into a stacked model (Wolpert, 1992), with ENLR serving as the learning algorithm. Classically, stacking is achieved using a disjunct mixing set, whereby the predictions of component models on the mixing set serve as the input features on which the stacking classifier is trained. Due to the limited amount of available data, this was achieved here by training partial component models in an inner 10x (distance-based) CV loop (**Figure 3**). Predictions of component models on all test sets were then concatenated into the training features of the stacking model. Predictions with the stacked model were made on the prediction output of the individual, full component models (XGB, ENLR, and SCM) (see **Supplementary Methods Section 2**).

## Model Evaluation

Component ML models as well as the stacking model were evaluated in the outer CV loop by predicting the MIC interpretive category (susceptible or resistant) on samples in the test set. Confusion matrices were summed up from outer CV folds. Performance of trained models was evaluated on the balanced accuracy (bACC) metric (Brodersen et al., 2010), as this metric allows evaluation of a model on imbalanced datasets. The bACC is furthermore related to the arithmetic mean of very major error (VME) and major error (ME), two performance criteria commonly applied to AST testing methods. Models created by the individual algorithms (XGB, ENLR, SCM), the

majority vote ensemble model and the stacking model were ranked by counting the number of other models achieving higher bACC on each organism/compound pair.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

LL, PM, SB, TR, and AP devised the study design. LL and PM wrote the code, performed experiments, and analyzed the resulting data. LL wrote the first draft of the manuscript. LL, PM, and SB wrote sections of the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by the Austrian Research Promotion Agency (FFG) (grants 866389, 874595, and 879570).

## ACKNOWLEDGMENTS

We thank Thomas Weinmaier for help with data retrieval, Michael Ante for fruitful discussion of the statistical analysis of results, and Anna Yuwen for critical reading of the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2021.610348/full#supplementary-material

## REFERENCES

AlQuraishi, M. (2019). ProteinNet: A standardized data set for machine learning of protein structure. *BMC Bioinf.* 20, 1–10. doi: 10.1186/s12859-019-2932-0

Aun, E., Brauer, A., Kisand, V., Tenson, T., and Remm, M. (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput. Biol.* 14, 1–17. doi: 10.1371/journal.pcbi.1006434

Aytan-Aktug, D., Clausen, P. T. L. C., Bortolaia, V., Aarestrup, F. M., and Lund, O. (2020). Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks. *mSystems* 5, 1–15. doi: 10.1128/msystems.00774-19

Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., Huang, B., et al. (2015). Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* 6, 1–14. doi: 10.1038/ncomms10063

Břinda, K., Callendrello, A., Cowley, L., Charalampous, T., Lee, R. S., MacFadden, D. R., et al. (2018). Lineage calling can identify antibiotic resistant clones within minutes. *bioRxiv* 403204, 455–464. doi: 10.1101/403204

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. *Proc. Int. Conf. Pattern Recognit.* 3121–3124. doi: 10.1109/ICPR.2010.764

Bunny, K. L., Hall, R. M., and Stokes, H. W. (1995). New mobile gene cassettes containing an aminoglycoside resistance gene, aacA7, and a chloramphenicol resistance gene, catB3, in an integron in pBWH301. *Antimicrob. Agents Chemother.* 39, 686–693. doi: 10.1128/AAC.39.3.686

Bush, K., and Jacoby, G. A. (2010). Updated functional classification of β-lactamases. *Antimicrob. Agents Chemother.* 54, 969–976. doi: 10.1128/AAC.01009-09

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. doi: 10.1145/2939672.2939785

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., et al. (2008). The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nat. Rev. Cancer* 8, 37–49. doi: 10.1038/nrc2294

Cox, G., Stogios, P. J., Savchenko, A., and Wright, G. D. (2015). Structural and molecular basis for resistance to aminoglycoside antibiotics by the adenylyltransferase ANT(2″)-Ia. *MBio* 6, 1–9. doi: 10.1128/mBio.02180-14

Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., et al. (2016). Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci. Rep.* 6, 1–12. doi: 10.1038/srep27930

Drlica, K., and Zhao, X. (1997). DNA gyrase, topoisomerase IV, and the 4-quinolones. *Microbiol. Mol. Biol. Rev.* 61, 377–392. doi: 10.1128/.61.3.377-392.1997

Drouin, A., Giguère, S., Déraspe, M., Marchand, M., Tyers, M., Loo, V. G., et al. (2016). Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics* 17, 1–15. doi: 10.1186/s12864-016-2889-6

Drouin, A., Letarte, G., Raymond, F., Marchand, M., Corbeil, J., and Laviolette, F. (2019). Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci. Rep.* 9, 1–13. doi: 10.1038/s41598-019-40561-2

FDA (2009). *Guidance for Industry and FDA Class II Special Controls Guidance Document : Antimicrobial Susceptibility Test ( AST ) Systems Preface Public Comment : Additional Copies*. Available at: https://www.fda.gov/medical-devices/guidance-documents-medical-devices-and-radiation-emitting-products/antimicrobial-susceptibility-test-ast-systems-class-ii-special-controls-guidance-industry-and-fda (Accessed December 7, 2020).

Ferreira, I., Beisken, S., Lueftinger, L., Weinmaier, T., Klein, M., Bacher, J., et al. (2020). Species identification and antibiotic resistance prediction by analysis of whole-genome sequence data by use of ARESdb: An analysis of isolates from the unyvero lower respiratory tract infection trial. *J. Clin. Microbiol.* 58, 1–11. doi: 10.1128/JCM.00273-20

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Software* 33, 1–22. doi: 10.1016/j.expneurol.2008.01.011

Hicks, A. L., Wheeler, N., Sánchez-Busó, L., Rakeman, J. L., Harris, S. R., and Grad, Y. H. (2019). Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. *PLoS Comput. Biol* 15, 1–21. doi: 10.1101/607127

Jacoby, G. A. (2005). Mechanisms of resistance to quinolones. *Clin. Infect. Dis.* 41, S120–S126. doi: 10.1086/428052

Karp, B. E., Tate, H., Plumblee, J. R., Dessai, U., Whichard, J. M., Thacker, E. L., et al. (2017). National Antimicrobial Resistance Monitoring System: Two Decades of Advancing Public Health Through Integrated Surveillance of Antimicrobial Resistance. *Foodborne Path. Dis.* 14, 545–557. doi: 10.1089/fpd.2017.2283

Kim, J., Greenberg, D. E., Pifer, R., Jiang, S., Xiao, G., Shelburne, S. A., et al. (2020). VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. *PLoS Comput. Biol.* 16, e1007511. doi: 10.1371/journal.pcbi.1007511

Kokot, M., Dlugosz, M., and Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33, 2759–2761. doi: 10.1093/bioinformatics/btx304

Kos, V. N., Déraspe, M., Mclaughlin, R. E., Whiteaker, J. D., Roy, P. H., Alm, R. A., et al. (2015). The Resistome of Pseudomonas aeruginosa in Relationship to Phenotypic Susceptibility. *Antimicrob. Agents Chemother.* 59, 427–436. doi: 10.1128/AAC.03954-14

Kuncheva, L.II, and Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Mach. Learn.* 51, 181–207. doi: 10.1049/ic:20010105

Lees, J. A., Galardini, M., Wheeler, N. E., Horsfield, S. T., and Parkhill, J. (2020). Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. *MBio* 11, 1–22. doi: 10.1128/mBio.01344-20

Mahfouz, N., Ferreira, I., Beisken, S., von Haeseler, A., and Posch, A. E. (2020). Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. *J. Antimicrob. Chemother.* 75, 3099–3108. doi: 10.1093/jac/dkaa257

Marchand, M., and Shawe-taylor, J. (2000). The Set Covering Machine. *J. Mach. Learn. Res.* 1, 723–746. doi: 10.1162/jmlr.2003.3.4-5.723

Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., and Parts, L. (2018). Precise prediction of antibiotic resistance in Escherichia coli from full genome sequences. *PLoS Comput. Biol.* 14, 2–17. doi: 10.1101/338194

Nguyen, M., Brettin, T., Long, S. W., Musser, J. M., Olsen, R. J., Olson, R. D., et al. (2018a). Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. *Sci. Rep.* 8, 1–11. doi: 10.1038/s41598-017-18972-w

Nguyen, M., Long, S. W., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., et al. (2018b). Using machine learning to predict antimicrobial minimum inhibitory concentrations and associated genomic features for nontyphoidal Salmonella. *J. Clin. Microbiol.* 57, 380782. doi: 10.1128/JCM.01260-18

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17, 1–14. doi: 10.1186/s13059-016-0997-x

O'Neill, J. (2016). Tackling Drug-Resistant Infections Globally. *J. Pharm. Anal.* 6, 71–79. doi: 10.1016/j.jpha.2015.11.005

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography (Cop)* 40, 913–929. doi: 10.1111/ecog.02881

Ruppert, D. (2004). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *J. Am. Stat. Assoc.* 99, 567–567. doi: 10.1198/jasa.2004.s339

Sharkey, L. K. R., and O'Neill, A. J. (2018). Antibiotic Resistance ABC-F Proteins: Bringing Target Protection into the Limelight. *ACS Infect. Dis.* 4, 239–246. doi: 10.1021/acsinfecdis.7b00251

Shaw, K. J., Rather, P. N., Hare, R. S., and Miller, G. H. (1993). Molecular genetics of aminoglycoside resistance genes and familial relationships of the aminoglycoside-modifying enzymes. *Microbiol. Rev.* 57, 138–163. doi: 10.1128/mmbr.57.1.138-163.1993

Sayers, E. W., Beck, J., Brister, J. R., Bolton, E. E., Canese, K., Comeau, D. C., et al (2020). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 48, D9–D16. doi: 10.1093/nar/gkz899

Strodthoff, N., Wagner, P., Wenzel, M., and Samek, W. (2019). Universal Deep Sequence Models for Protein Classification. *bioRxiv* 704874, 1–11. doi: 10.1101/704874

Tabe-Bordbar, S., Emad, A., Zhao, S. D., and Sinha, S. (2018). A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. *Sci. Rep.* 8, 1–11. doi: 10.1038/s41598-018-24937-4

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2019). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods Ecol. Evol.* 10, 225–232. doi: 10.1111/2041-210X.13107

Valizadehaslani, T., Zhao, Z., Sokhansanj, B. A., and Rosen, G. L. (2020). Amino acid K-mer feature extraction for quantitative antimicrobial resistance (AMR) prediction by machine learning and model interpretation for biological insights. *Biol. (Basel)* 9, 1–92. doi: 10.3390/biology9110365

Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., et al. (2016). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* 45, 535–542. doi: 10.1093/nar/gkw1017

Wayne, P. (2019). *Performance standards for antimicrobial susceptibility testing. 29th ed. CLSI supplement M100* (Wayne, PA: Clinical and Laboratory Standards Institute).

Wolpert, D. (1992). Stacked Generalization. *Neural Networks* 5, 241–259. doi: 10.1016/S0893-6080(05)80023-1

**Conflict of Interest:** LL, PM, SB, and AP are employed by Ares Genetics GmbH.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2021 Lüftinger, Májek, Beisken, Rattei and Posch. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Digital Health

PERSPECTIVE
published: 16 February 2021
doi: 10.3389/fdgth.2020.583390

# How to Develop and Implement a Computerized Decision Support System Integrated for Antimicrobial Stewardship? Experiences From Two Swiss Hospital Systems

*Gaud Catho[1,2]*, Nicolo S. Centemero[3], Brigitte Waldispühl Suter[3], Nathalie Vernaz[2,4], Javier Portela[5], Serge Da Silva[5], Roberta Valotti[6], Valentina Coray[3], Francesco Pagnamenta[3], Alice Ranzani[1], Marie-Françoise Piuz[5], Luigia Elzi[7], Rodolphe Meyer[5], Enos Bernasconi[6], Benedikt D. Huttner[1,2] and the COMPASS Study Group*

[1] Division of Infectious Diseases, Geneva University Hospital, Geneva, Switzerland, [2] Faculty of Medicine, University of Geneva, Geneva, Switzerland, [3] Division of Clinical Informatics, Ente Ospedaliero Cantonale, Bellinzona, Switzerland, [4] Medical Direction, Geneva University Hospital, Geneva, Switzerland, [5] Division of Infectious Diseases, Ospedale Regionale di Lugano, Ente Ospedaliero Cantonale, Lugano, Switzerland, [6] Division of Infectious Diseases, Ospedale San Giovanni, Ente Ospedaliero Cantonale, Bellinzona, Switzerland, [7] Division of Informatics, Geneva University Hospital, Geneva, Switzerland

OPEN ACCESS

**Edited by:**
*Belén Rodriguez-Sanchez, Gregorio Marañón Hospital, Spain*

**Reviewed by:**
*Jian Guo, RIKEN Center for Computational Science, Japan*
*Xia Jing, Clemson University, United States*

***Correspondence:**
*Gaud Catho*
*gaud.catho@hcuge.ch*

**Specialty section:**
*This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health*

**Received:** *14 July 2020*
**Accepted:** *25 November 2020*
**Published:** *16 February 2021*

**Citation:**
*Catho G, Centemero NS, Waldispühl Suter B, Vernaz N, Portela J, Da Silva S, Valotti R, Coray V, Pagnamenta F, Ranzani A, Piuz M-F, Elzi L, Meyer R, Bernasconi E, Huttner BD and the COMPASS Study Group (2021) How to Develop and Implement a Computerized Decision Support System Integrated for Antimicrobial Stewardship? Experiences From Two Swiss Hospital Systems. Front. Digit. Health 2:583390. doi: 10.3389/fdgth.2020.583390*

**Background:** Computerized decision support systems (CDSS) provide new opportunities for automating antimicrobial stewardship (AMS) interventions and integrating them in routine healthcare. CDSS are recommended as part of AMS programs by international guidelines but few have been implemented so far. In the context of the publicly funded COMPuterized Antibiotic Stewardship Study (COMPASS), we developed and implemented two CDSSs for antimicrobial prescriptions integrated into the in-house electronic health records of two public hospitals in Switzerland. Developing and implementing such systems was a unique opportunity for learning during which we faced several challenges. In this narrative review we describe key lessons learned.

**Recommendations:** (1) During the initial planning and development stage, start by drafting the CDSS as an algorithm and use a standardized format to communicate clearly the desired functionalities of the tool to all stakeholders. (2) Set up a multidisciplinary team bringing together Information Technologies (IT) specialists with development expertise, clinicians familiar with "real-life" processes in the wards and if possible, involve collaborators having knowledge in both areas. (3) When designing the CDSS, make the underlying decision-making process transparent for physicians and start simple and make sure to find the right balance between force and persuasion to ensure adoption by end-users. (4) Correctly assess the clinical and economic impact of your tool, therefore try to use standardized terminologies and limit the use of free text for analysis purpose. (5) At

the implementation stage, plan usability testing early, develop an appropriate training plan suitable to end users' skills and time-constraints and think ahead of additional challenges related to the study design that may occur (such as a cluster randomized trial). Stay also tuned to react quickly during the intervention phase. (6) Finally, during the assessment stage plan ahead maintenance, adaptation and related financial challenges and stay connected with institutional partners to leverage potential synergies with other informatics projects.

**Keywords: antimicrobial stewardship, implementation, digital health, usability testing, cluster randomized controlled trial, multidisciplinary, user training, computerized decision support system**

## INTRODUCTION

Antimicrobial resistance (AMR) remains one of the major global public health threats of the early twenty-first century and although detailed data are currently lacking, one that is potentially exacerbated by the current COVID-19 pandemic. Similar to SARS-CoV-2, albeit at a much slower pace, AMR is a pandemic with new multidrug resistant clones of pathogens continuing to emerge and spread globally, threatening our ability to treat common infectious diseases, ultimately resulting in prolonged illness, disability, and even increased risk of death (1). Antibiotic use is a key driver for the spread of AMR. While antibiotics have dramatically changed the prognosis of many common severe bacterial infections, they are among the most misused and overused medicines worldwide.

Antimicrobial stewardship (AMS) programs use different interventions to influence the behavior of prescribers toward a more rational and appropriate use of antimicrobials to improve patient care and preserve this resource for future patients and generations (2, 3).

Over the last decades, information technologies (IT) have become essential components of modern medicine and significantly impacted the delivery of health care. Electronic health records (EHR) now usually incorporate computerized physician order entry (CPOE) systems that not only assure trackability and documentation of prescriptions but may also enable computerized decision support systems (CDSS) that support physicians and other healthcare workers (HCWs) to optimize their decision-making.

Taking numerous, often complex and sometimes high-impact (for the patient and society) decisions under time pressure is part of the daily routine of HCWs around the world. Those decisions are often made *ad hoc* during patient contact, ward rounds or multidisciplinary meetings based on the medical knowledge and patient information available and accessible to the HCWs at the time of the decision. CDSSs offer the possibility to complement the information available for the HCWs by patient-specific and updated evidence-based recommendations at the point of care. CDSS have been shown to reduce medical errors, increase adherence to guidelines and ultimately increase patient safety (4, 5).

About 30–50% of patients will receive antimicrobials during their hospital stay (6) and those prescriptions are usually performed by physicians without specific training in infectious diseases and often only rudimentary knowledge about the appropriate use of antimicrobials. Furthermore, the epidemiology of disease-causing microbes is quickly evolving and varies among settings, making it challenging for non-infectious diseases (ID) specialists to stay updated when changing locations or when new versions of guidelines are released.

The time and economic constraints of modern healthcare delivery make it impossible to have every antimicrobial prescription assessed by ID experts. As part of AMS programs, post-prescription review of antimicrobial prescriptions by experts has been shown to improve antibiotic prescribing but is resource intensive and cannot be generalized (7). CDSS directly integrated into EHRs have the potential to promote the appropriate use of antimicrobials by providing prescribers with relevant real-time patient, alerts and recommendations when the prescribing decision is taken, without need for intervention by a specialist.

The COMPASS tool is a CDSS developed in the context of the COMPASS trial, a cluster-randomized, parallel-arm, open-labeled, superiority trial that aim to assess the effectiveness of a multi-modal computerized antimicrobial stewardship intervention (8). The COMPASS CDSS was developed between 2017 and 2018 implemented in 2018 in two hospital organizations in Switzerland: Geneva University Hospitals (HUG) and Ticino Regional Hospitals (EOC). The EHRs in both hospitals are in-house systems, which offer the flexibility to develop new components such as CDSSs integrated directly into CPOE.

## OBJECTIVE

In this article we aim to describe the process of developing a CDSS for the purpose of AMS from the point of view of clinician-investigators. We report some of the challenges we encountered and share the lessons we learned (**Table 1**). In the first part we describe issues related to the planning and development stages, in the second part we present issues related to implementation and evaluation.

## MAIN SECTION

The COMPASS CDSS provides guidance to physicians for in-patient clinical management. When prescribing antimicrobials

**TABLE 1 |** Key messages when designing and implementing your CDSS for antimicrobial prescriptions.

| |
|---|
| **Planning and development** |
| Draft the CDSS as an algorithm and use a standardized format |
| Set-up a multidisciplinary team bringing together IT specialists with development expertise, clinicians familiar with "real-life" processes in the wards and communicate clearly with members of the project and related stakeholders |
| Make the underlying decision-making process transparent for physicians and start simple |
| Find the right balance between force and persuasion |
| Beware of the planning fallacy |
| **Implementation** |
| Plan usability testing early and regularly in the developing process |
| Think ahead of additional challenges related to study design and stay tuned to react quickly during the intervention phase |
| Plan training appropriately |
| **Assessment and adaptation** |
| Plan ahead maintenance, adaptation and related financial challenges |
| Potentialize synergies with other IT projects |

on inpatient wards, physicians must choose the indication for each antimicrobial; thereafter they are provided indication-specific treatment recommendations based on local guidelines. After 3 days of therapy physicians receive a prompt for a self-guided evaluation of the prescription. As part of the intervention of the COMPASS study, physicians working in wards where the CDSS was implemented also received quarterly feed-back on qualitative antibiotic use data from the CDSS.

## Planning and Development Stage

### Message 1: Draft the CDSS as an Algorithm and Use a Standardized Format

The COMPASS project was an investigator-initiated project funded by the Swiss National Science Foundation in the context of a national research program on antimicrobial resistance (9).

The clinical investigators (BH, GC, EB) had some basic knowledge of informatics but no IT background. One key challenge when designing a CDSS is that clinicians and informaticians may not necessarily share the same language and concepts. A crucial first step when designing a CDSS is to develop a concept of the tools desired functionalities and algorithms and share it early on with software developers to assess feasibility and necessary modifications. We used simple algorithms and described them in a schema (created in OmniGraffle, The Omnigroup, Seattle, United States) providing a clear and concise outline of the functionality we expected. A standardized format such as Business Process Model and Notation (BPMN) (http://www.bpmn.org/) can be used for this purpose. These formats are particularly useful to represent workflow and rules of the clinical decision support tools (**Figure 1A**). They describe procedures using a graphical notation and give the ability to communicate these procedures in a standardized manner. They are also useful to document the processes for future reference.

A further helpful tool are Digital Accelerator Kits (DAK) developed by the World Health Organization (WHO) to translate narrative guidelines into a standardized format that can be more easily digitalized and integrated into decision support systems. DAK's consist of the standardized documentation of the foundational components of digital client records, including common workflows, core data elements, decision-support algorithms and scheduling logic, metrics and reporting indicators. DAK have been designed to ensure WHO's evidence-based guideline content is accurately reflected in the digital systems that countries are adopting. Using standardized graphical presentation such as components proposed by the DAK or BPMN format help to make the process transparent and understandable by clinicians and software developers. It also makes the CDSS readable by stakeholders not involved in the initial development stage in case of further evolution or adaptation of the CDSS (10).

**Figure 1B** presents the initial algorithm of the COMPASS tool as it was conceived by the PI of the project and **Figure 1A** the final algorithm drawn in collaboration with business modelers and software developers at the end of the development stage. Retrospectively, we think that we could have saved time and avoided misunderstandings by using a standardized format and by involving someone with skills in graphical presentation of informatics processes early in the development stage.

### Message 2: Set-Up a Multidisciplinary Team Bringing Together Information Technology (IT) Specialists With Development Expertise, Clinicians Familiar With "Real-Life" Processes in the Wards and Communicate Clearly With Members of the Project and Related Stakeholders

Many CDSS are developed by software developers without early involvement of clinicians. Clinicians have an intimate understanding of healthcare delivery, having spent thousands of hours in clinical settings in training and practice. Having a good understanding of how the EHR works from the end-user perspective and of the exact prescribing workflow is key when designing a CDSS that targets prescribing behaviors. It allows an effective validation feedback loop during each development step and makes CDSS fit with clinician workflow. This point is strongly associated with a decision support system's ability to improve clinical practice (11). Each step can be tested and validated by users that perceive real-life problems that might emerge. On the other hand, clinicians lack the IT background to understand the feasibility and time necessary for implementing certain functionalities.

Our COMPASS project was developed at two different sites (Ticino and Geneva) with two independent teams composed by IT with development expertise and clinical researchers (ID

physicians). Both tools were based on the same algorithms and communication between the two teams during the development occurred frequently. In one study site, three members of the IT team had also a background as medical doctor and pharmacist. They were playing a key role by being at the interface between clinicians and back-end developers. We realized that sometimes there were misunderstandings and communication problems due to the different backgrounds of the involved experts. We therefore strongly recommend involving someone in the project early on who has expertise both in clinical medicine and IT and who can understand both languages and "translate" between different experts. The recently published findings of a Delphi panel highlighted the added value of hybrid positions that blend responsibilities, knowledge, and experience in clinical quality, patient safety, and informatics (12).

During the development of our CDSS, one of the study sites was in the process of establishing a workflow for validation of informatics projects. It therefore happened that decisions taken by the investigators together with the IT-team were later put into question by a different entity. A key lesson we learned is that assuring frequent communication, identifying and implicating all relevant partners and clearly establishing tasks and responsibilities for each partner from the beginning is key for the successful development of such a complex project.

## Message 3: Make the Underlying Decision-Making Process Transparent for Physicians and Start Simple

During the qualitative study that we conducted before implementation of the COMPASS trial (13), we found that transparency about how the CDSS makes output decisions is a key factor for CDSS acceptance by physicians. Physicians are reluctant to trust a "black-box" system if they cannot assess the pathway that led them from the diagnosis to the proposition made. Physicians who understand what the computer recommendations are based on more willing to accept it (14). Our system was based on a relatively simple algorithm (recommendations based on the indication selected by the prescriber) that could fit on a single screen. All our pre-existing antimicrobials guidelines were translated into the CDSS. When several options exist for a specific indication (based on various susceptibility profiles for the same pathogen), all the propositions are displayed to clinicians with the rules that condition the choice mentioned as free text in specific boxes. The physician decides which options to choose based on the characteristics of his/her patient (such as previous microbiologic results). Studies have shown that simple interventions often work best (15).

We also found that providing the sources of the recommendations matters for adoption. Our CDSS was based on local recommendations established by the infectious diseases department and already available through a booklet or PDF. Most users were therefore already familiar with propositions of

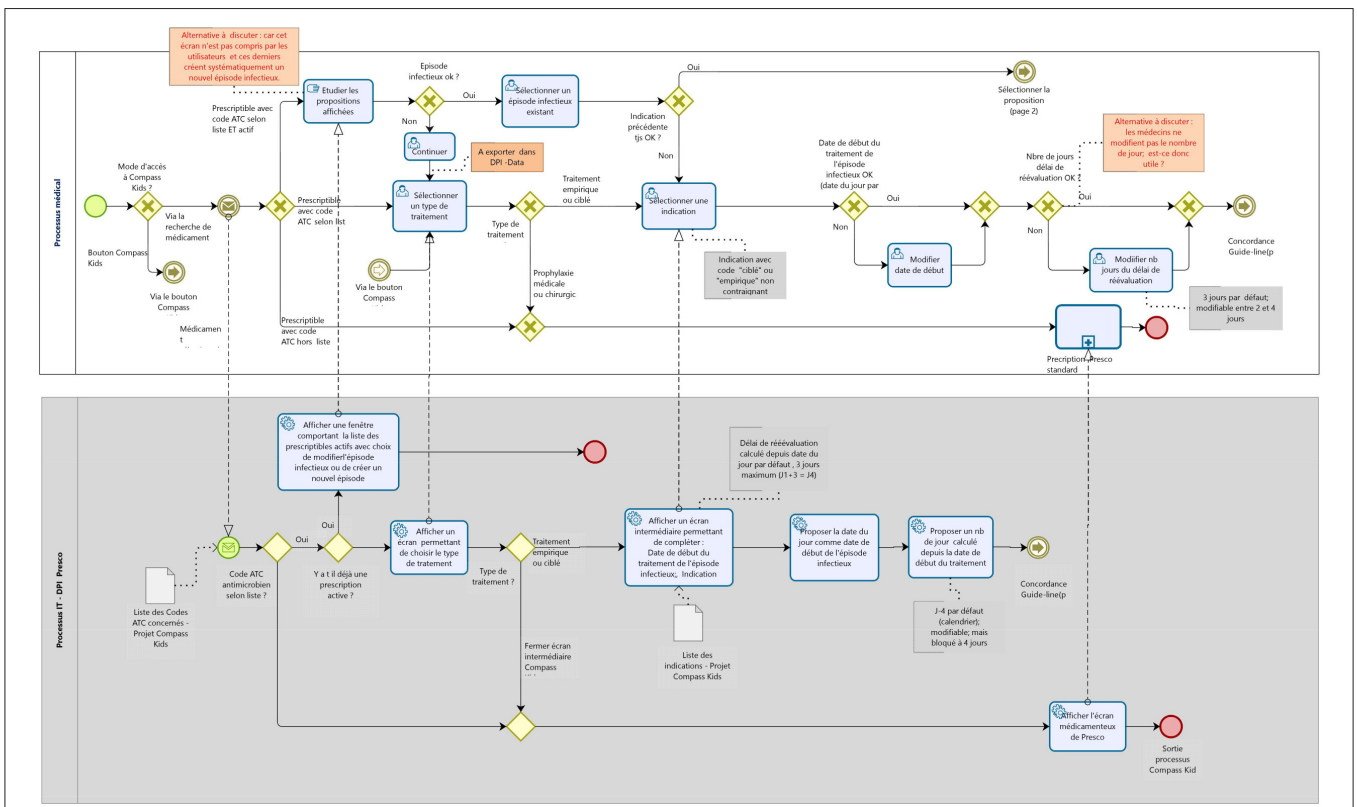



**FIGURE 1 |** Continued

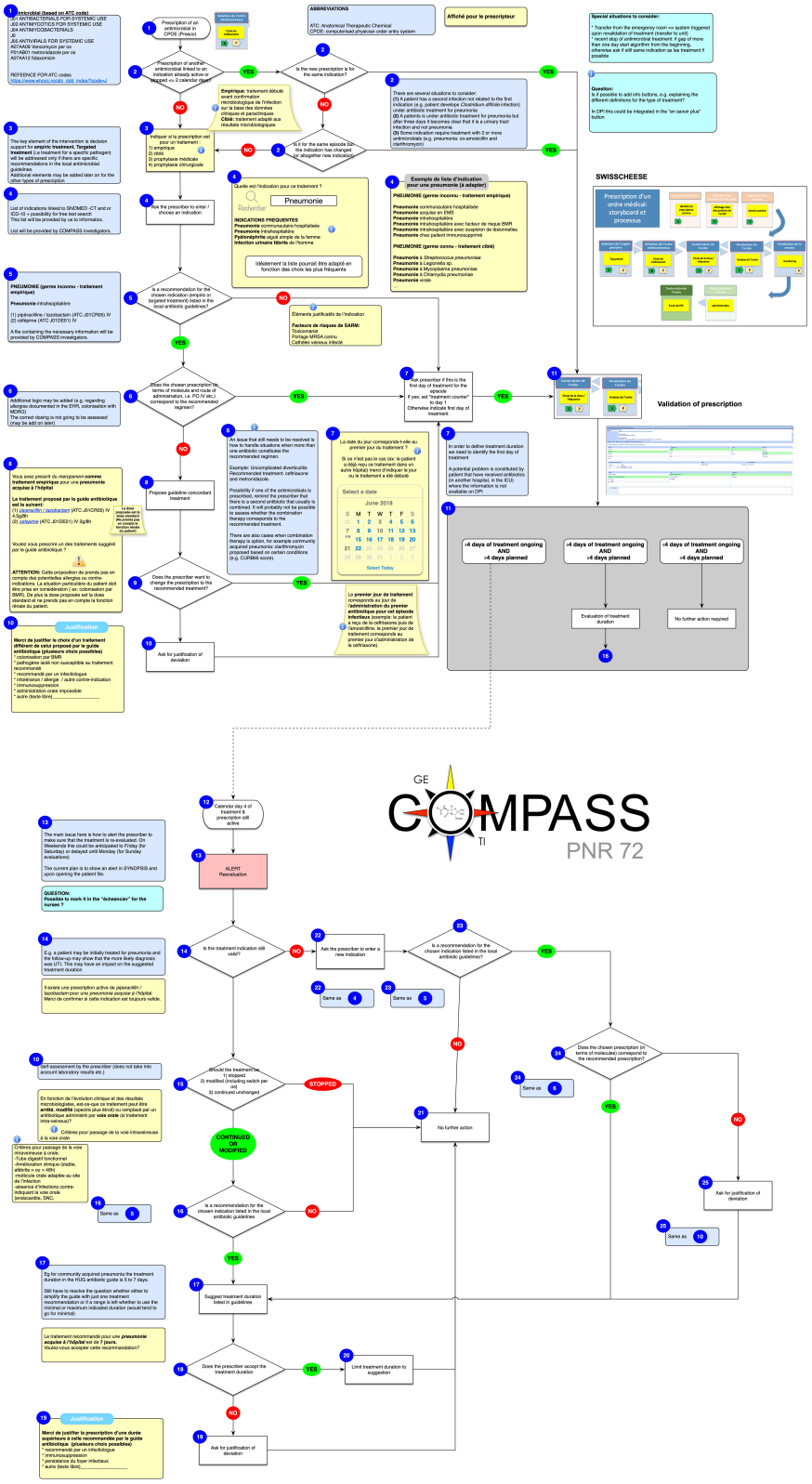



**FIGURE 1 | (A)** Workflow and rules of the COMPASS algorithm drawn in collaboration with business modelers. **(B)** Workflow and rules of the COMPASS algorithm initially drawn by the PI.

the CDSS and we made clear in the CDSS and in the training session we provided that recommendations in the booklet and the CDSS were similar, with the only exception that the CDSS can be updated more frequently. A link to the PDF was also created in the CDSS.

When designing your CDSS we recommend to keep the underlying processes simple and visible for end-users and make clear for them where the content is coming from.

## Message 4: Find the Right Balance Between Force and Persuasion

To have a CDSS actively used by prescribers, and therefore being able to assess its impact, it is crucial to choose an appropriate trigger or entry point. In our case we chose as trigger the entry of an order for an antimicrobial in the CPOE. This trigger integrates well into the clinician workflow since the physicians were receiving recommendations immediately after having taken the decision to prescribe an antimicrobial. A recent study reviewing the rule-based clinical decision support content of a large integrated delivery network found that "order entry" trigger accounted for 94% of all triggers for the studied clinical rules and 38% of all clinical rule types (16).

In COMPASS each time a physician ordered an antimicrobial (from a list of selected antimicrobials based on ATC codes; HIV medicines were e.g., excluded), he or she was forced to use the CDSS.

The cluster-randomized design makes this rule more complex: the initial CDSS development in Geneva hospital system did not allow automatic triggering of the CDSS for patients transferred to an intervention unit from a non-intervention unit who had already been prescribed antimicrobials in that unit. In order for the CDSS to be used in these instances, physicians had to "manually" stop antimicrobials and prescribe them again through the CDSS on a voluntary basis. In this context a significant proportion of antimicrobial prescriptions were not made through the CDSS since many patients arrived in the intervention units with antimicrobials prescribed in the emergency room and physicians perceived the stopping/re-prescribing as a loss of time. It is noteworthy to mention that this problem was "artificial" in the context of the trial. Indeed, if the CDSS were to be implemented in every unit of the hospital, including the emergency room, all antimicrobials would be prescribed through the CDSS from the beginning and stopping/re-prescribing antimicrobials would not be necessary.

This initial low uptake of the CDSS threatened the validity of the study since significant underuse of the tool would have not allowed to assess the effectiveness of the tool itself (a tool that is not used cannot be expected to have an effect). An additional development was performed to address this problem few months after the initial launch and the use of the CDSS for the patients in an intervention unit and already receiving antimicrobials became mandatory. In comparison, in Ticino this feature was implemented from the beginning of the implementation.

On the other hand, for safety reasons, we decided to not make the self-guided re-evaluation of the prescription mandatory. Instead of automatically stopping a prescription not evaluated after 3 days, prescriptions were presented in a gray banner and marked as "to be re-evaluated" (**Figure 2**). This display persisted until the reevaluation task was completed but had no direct impact on the prescription, meaning that without any action the prescription would continue as originally planned. We observed here the limits of a persuasive system as the action of reevaluation was poorly performed by end-users.

The right balance between persuasive and restrictive strategies is difficult to achieve. By being too restrictive and forcing or blocking the prescribers, one risks limiting their autonomy too much and thus decreasing the acceptability of the system (17). By being too permissive and providing only suggestions, the resistance to change may result in prescribers not using the system, precluding any chance for an impact. We recommend to carefully select which part of your intervention needs to be mandatory to be able to correctly assess its impact.

Published data on strategies to encourage prescribers to perform self-guided review of antibiotics regimens report that these strategies should include persuasive or enforced prompting. Without such mechanisms, these interventions are likely to have minimal impact (3). A recent review identified factors associated with the successful implementation of persuasive interventions (18). The authors report that provider education should be part of any multimodal intervention that includes a persuasive strategy. Interestingly, patient integration and empowerment was also associated with successful implementation. We could imagine that this prompt for reevaluation might trigger a discussion with the patient or his family on the antibiotic strategy.

The override of reevaluation alerts by physicians might have several other explanations besides the facultative aspect, such as poor design of the response mechanism. The perception that the system is merely giving an assessment ("*your prescription has to be reevaluated*") without recommending an action and providing a convenient way to either carry out or disregard has been described as an inneffective way to change behavior (19). More complex decision rules such as "*your patient is already receiving oral drug, to switch to an oral antibiotic click here*" or "*your patient has been treated more than 5 days for a pneumonia, to stop it click here*" might have received higher acceptance.

When designing a multimodal intervention, we recommend combining restrictive and persuasive strategies associated with prescriber education and involvement of patients and families.

## Message 5: Beware of the Planning Fallacy

We have been confronted during the whole of the project to what is commonly called the "planning fallacy," i.e., the tendency for humans to "underestimate the time it will take to complete a future task, despite knowledge that previous

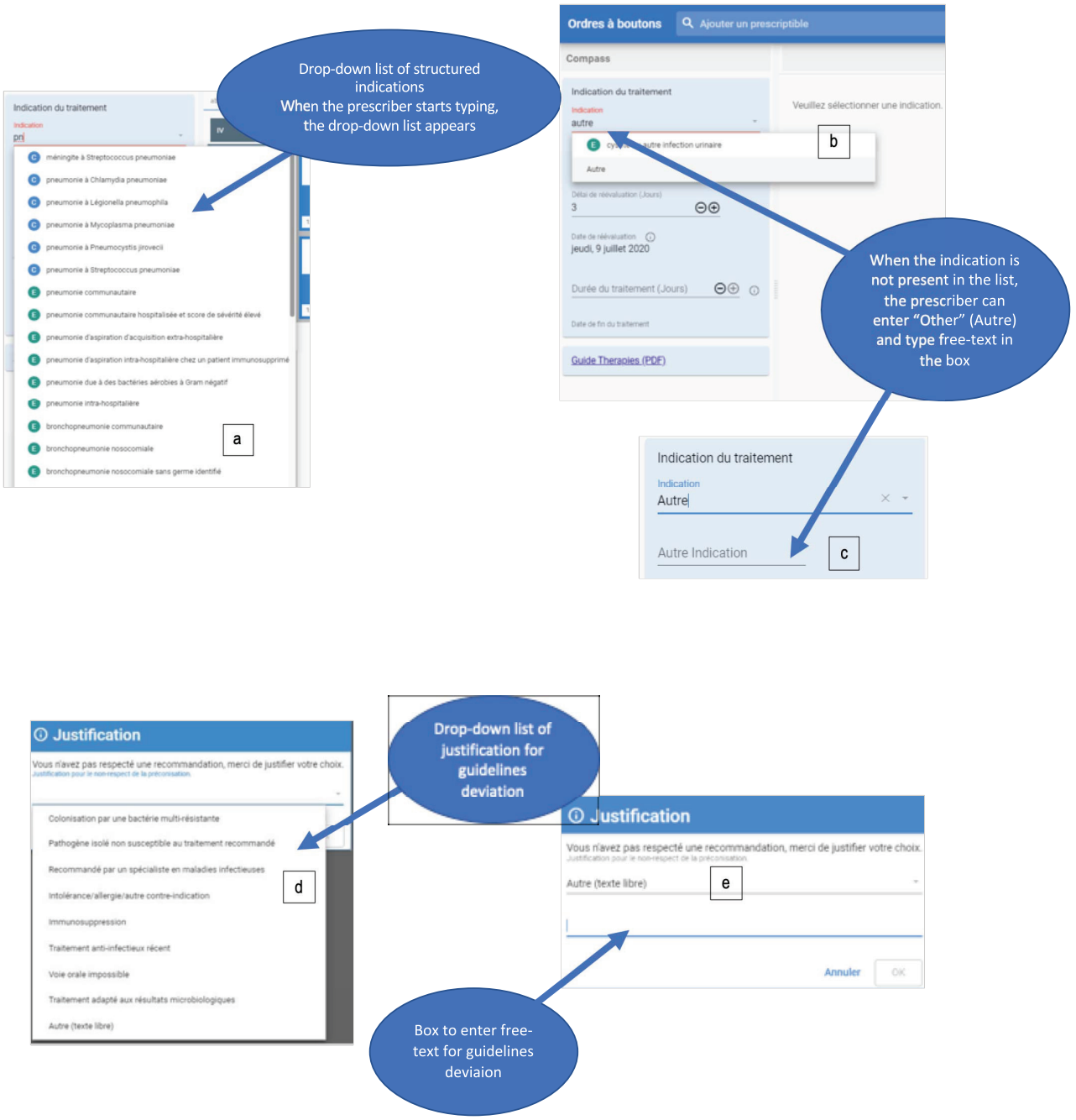



**FIGURE 2 |** Screenshots of the CDSS. **(A)** Dropdown list of indications, **(B)** selection of free-text indication by selecting "Autre" ("Other") in the dropdown list and **(C)** entering free-text. **(D)** List of justification provided for guideline rejection, **(E)** by selecting "Autre" (Other), the specific box for free-text appears.

tasks have generally taken longer than planned" (20). The planning fallacy can have deleterious implications for any project, and it was one of the major obstacles we encountered. For our COMPASS project, at initial timeline, the CDSS would have been launched in April 2018, it was finally implemented in September 2018 in Ticino and December 2018 in Geneva.

With hindsight the initial timeline was clearly overly optimistic. In addition to some "naïve" assumptions by the investigators, the pressure from funding agencies to rapidly implement research projects and obtain results in time, and limited budget certainly also played a role in this initial timeline. On the other hand, the informatics departments of our hospitals (both in Ticino and Geneva) were confronted with many

competing projects and maintenance tasks that could not be post-poned, a reality of which the clinical investigators were not necessarily aware.

Identifying the potential cause of delays early on and, discussing with experts from all implicated domains can help to mitigate the planning fallacy. A qualitative study identified 15 potential socio-technical challenges leading to delays in CPOE and CDSS implementation (21). They differentiate unintended delays from tactical internal delays. Tactical delays are due to tactical decisions taken to enhance longer-term adoption and optimized use of the system.

If some strategies can mitigate part of the project delays (such as detailed planning, acquiring better knowledge of systems, and stepwise implementation strategies), many other delays are unavoidable, and this should be kept in mind when starting medical informatics projects. Simplistic and unrealistic assumptions at the early stages of a project are unhelpful in making decisions for planning of medical technology projects and lead to frustration on all sides. Adequate time and effort must be spent at early stages of a project to capture the needs of short- and long-term users, system benefits and implementation strategies.

Our recommendation would be to make a clear schedule to outline every step of the project with regular assessment of tasks and deliverables, to ensure that everyone is on the same page about the requirements and to make realistic assumptions about resource availability and deadlines.

## Message 6: Use Standardized Terminologies and Limit the Use of Free-Text

There is a trade-off between the use of structured information and free text that users can enter into the CDSS. From an analysis perspective, free texts will induce considerable additional workload to reclassify information into structured terminology. The different experience in the two hospital systems participating in COMPASS nicely illustrates the behaviors of end-users with regard to this aspect. In the COMPASS CDSS the initial step for the prescriber is to select an indication for the antimicrobial he/she prescribes (**Figure 2**). In one center, to enter free-text, the prescriber has to type "Other," then a new box will appear where he/she can type a free-text indication. Due to this "trick" that makes entering free-text difficult to find, we ended up with very few free-text indications entered in one center. In the other center it was much easier to enter free text, resulting in a much higher proportion of unstructured indications (when reviewing the indications, a high percentage would have been available as structured information, but the end-user was unable to find it or did not make the effort to find it).

On the opposite, in the center where the free text indication was hidden, the list of justifications for deviation from recommendations contains "other" directly visible in the list below the 6 other propositions (**Figure 2**). The amount of free text justifications was considerably larger than free-text indications. This illustrated that users will generally favor free-text when available because it costs them less effort than to search for the indication from a predefined list.

We recommend limiting the possibility to enter free text when designing your own CDSS, but finding the right concept from predefined lists should be made as easy as possible. One hospital (Geneva) maintains a list of ∼50,000 medical terms coded and linked to international terminology such as ICD-10 (22). Due to communication issues between the different databases, implementing this list in our own system was not possible. We selected infectious diseases terms from the list and indexed all the possible alias for each term (example **Figure 2**). The final list contains more than 500 indications (including many aliases) in Geneva and about 200 in Ticino. Results from a survey conducted among users shows that among 10 users who entered comments on indications, 6 of them complained that there was not enough indications or the indication they were looking for was not present in the list; on the other hand 4 users expressed that there were too many indications with too many aliases and that they struggled to find the proper one. The fact that nearly as many end-users complained about too many as about too few indications suggests that we probably found a good balance.

We recommend limiting use of free-text but also adapting your tool over time (e.g., here include in the list of indications that were not found and provided as free-text).

## Message 7: Plan Usability Testing Early and Regularly in the Developing Process

Usability refers to the ease of use of an interface, defined in part by learnability, efficiency, memorability, satisfaction, and potential for errors. Usability of an informatic product is crucial for adoption by end-users (23, 24). Usability testing is part of what is now commonly called User experience (UX) and refers to the methods for improving ease-of-use during the design process, generally testing it with representative users. Typically, during a test, participants will try to complete typical tasks while observers watch, listen and take notes ("think aloud" method) (25). The goal is to identify any usability problems, collect qualitative and quantitative data and determine the participant's satisfaction with the product.

Effective CDSS are often the product of an interactive design process based on usability evaluation and redesign. While this process might be perceived as time-consuming and laborious, it may detect significant problems and considerably increase user's satisfaction with your system. Usability technique does not necessarily need complex methods or formal lab equipment, but it needs to be planned ahead and budgeted. Ideally, usability testing should be performed early in the design process and throughout the development cycle (15).

At the start of our project, usability testing of in-house informatic products was not routinely implemented in our institutions yet and we had the feeling that usability testing of our tool could only be performed once the tool was almost ready. However, testing your product too late in the development process might lead you to a point when corrections will be much more costly. Furthermore, as clinicians involved in developing the tool, we became used to its imperfections and the "tricks" to circumvent them, therefore testing it by ourselves became somewhat misleading and did not necessarily reflect "real-life."

When our tool was finally ready to be launched, delay in development and the time pressure in the context of the research study made us bypass an extensive usability testing. Before implementation we performed tests with users to validate that they were able to use the tool in classic scenarios, but no formal testing with "think aloud" methods or deep analyses of problems testers encountered were assessed. At this stage any changes in the design or algorithm would have cost us some additional months of delays. In hindsight, we feel that important, although not necessarily time-consuming changes could have had significant impact in adoption by end-users and that it would have been worth to invest more time and resources for usability testing. For example, the box to allow the prescriber to add a medication to the proposed regimen was not visible enough in the first version of the tool. Only few months after the launch, the lay-out of this box was slightly modified to be find more easily by prescribers. This type of "mistakes" could have been detected through basic usability testing and corrected from the beginning.

We strongly recommend to carefully plan and budget usability testing when designing your own CDSS.

## Implementation Stage

### Message 8: Think Ahead of Additional Challenges Related to Study Design and Stay Tuned to React Quickly During the Intervention Phase

Cluster randomized trials are considered the best study design to assess AMS interventions for several reasons (26): they reduce the risk of contamination of the intervention, may enhance compliance to the intervention within the cluster and allow to assess specific outcomes such as antimicrobial resistance at the cluster level.

Implementing a computerized decision support system in the context of a cluster randomized trial adds additional challenges to the well-described logistical and financial challenges inherent to this type of design.

The CDSS access has to be restricted to physicians and patients in specific units. It means that, when a patient is transferred from an intervention to a control unit, information related to the CDSS (indication for antimicrobial prescriptions) has to be hidden. To avoid prescriber-fatigue, ideally this information should be kept and reappear in case of patient re-transfer in the other direction. It appears for example that in case of a short stay in operatory room of a patient from an intervention ward, all the indication data were lost, and prescribers had to re-enter data again (and then re-prescribe the drugs) when the patient was back. This type of event potentially creates considerable frustration for the prescribers. They need to be think ahead, continuously monitored and quickly corrected.

When implementing the CDSS in the context of a study, small mistakes at the implementation phase can compromise the entire study. During the study period, you need to be particularly alert of small disruptions not planned beforehand and maintain a constant communication with developers and end-users to detect problems early on and react fast with appropriate corrective actions.

### Message 9: Plan Training Appropriately

Regardless of its self-learnability, all new systems have a learning period, and so baseline evaluations of users' technological competence may be appropriate. Further training can be provided to facilitate full use of CDSS capabilities or more explicit guidance incorporated into the CDSS' recommendations themselves. This information could be implemented as info buttons to be non-disruptive. There is again a trade-off between too much info buttons that will disturb end-users and not enough which will not allow those who did not receive specific training to fully exploit potentialities of the CDSS. As mentioned in a recent systematic review, research is needed to investigate user experience improvements to increase info button use and effectiveness (27).

We observed that our system was not fully intuitive to allow a comprehensive use without additional training. In-person training session and on-ward in-person support was performed during the first weeks in both study sites. In Ticino only, in-person training sessions were mandatory. In Geneva, we created extensive add-on training materials such as an intranet website, frequently ask questions documents, and small demo-videos. Nevertheless, we had the feeling that very few people made the effort to look at this information. People are used to very performant electronic tools e.g., smartphones, tablets in their daily life and their level of expectations toward informatic products is high. They want a tool intuitive enough not to require additional efforts to become familiar with. In this sense, usability testing mentioned above can make big differences through small changes.

## Assessment Stage and Adaptation

### Message 10: Plan Ahead Maintenance, Adaptation, and Related Financial Challenges

Maintenance and continuous adaptation of CDSS are others challenges that can be frequently neglected at the initial stage (16). Maintenance is crucial for two main reasons: (1) the content of the clinical rules needs to be regularly adapted to the underlying evidence-based knowledge, which is itself evolving fast, (2) technical adjustments might be necessary due to updates or new functionalities in the EHR or CPOE.

To regularly update content, we recommend building a system that allows a certain degree of autonomy for clinicians. In our case, a web-based platform was designed that allows modifying order sets by clinicians in charge of the project. Any modification of the content of the local guidelines is under the responsibility of the infectious diseases division and validated by the team in charge of local guidelines before dissemination. This platform was integrated into the production version of the EHR and therefore any changes could be quickly released in production without a long and frustrating validation process. Nevertheless, to keep up with the pace of changes in medical knowledge and local guidelines requires time.

Regarding new functionalities, alerts for drug-drug interactions and renal dosing adaptations are planned to be implemented in the electronic prescribing systems of both institutions after the launch of the COMPASS study. These new features required a very careful evaluation that new

functionalities will also be effective when prescribing through the COMPASS system. Users are becoming quickly familiar with these additional features and not having them available can be frustrating.

Nevertheless, even if solutions for maintenance and adaptation can be found, their costs are another challenge. In our case, the development of CDSS was part of a research project financed by the Swiss National Science Foundation. It means that the budget ends once the research project is over. Proving cost-effectiveness or improvement of quality of cares is crucial for convincing institutional leaders to keep financing maintenance and further developments.

### Message 11: Potentialize Synergies With Other Informatic Projects

The informatics development in the hospital context should be seen in a broad perspective. Subcomponents of our COMPASS CDSS have already been re-used for other informatic development for example during the COVID-19 crisis to create local multi-component guidelines and for a similar tool targeting prescriptions of antimicrobials to hospitalized children in the pediatric hospital in Geneva. Synergies can also be created between other informatic projects targeting medical prescriptions. By being proactive and aware of concomitant informatics projects developed in your own institution, you can create bridges and leverage the development performed for other projects.

## CONCLUSION

Developing our own CDSS for antimicrobial stewardship was a very exciting but also challenging experience. Having our own in-house Electronic Health Record offered us this (increasingly rare) opportunity to build a CDSS integraded into the electronic prescribing of our hospitals. Nowadays, big commercial EHRs are replacing local in-house systems which limit the possibilities do add new functionalities adapted to local needs and practices. Developing a CDSS in collaboration with IT teams was a multifaceted experience with some unforeseen challenges.

Assessing the real impact of those tools is key and the literature is clearly lacking so far. We are looking forward sharing the results of the cluster-randomized trial.

## COMPASS STUDY GROUP

Carlo Balmelli, Stefano Bruni, Magali Despond, Emmanuel Durand, Laurent Kaiser, Damien Grauser, Stephan Harbarth, Christophe Marti, Virgini Prendki and Jerome Stirnemann.

## DATA AVAILABILITY STATEMENT

The original contributions generated in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

GC and BH: conceptualization. GC: writing—original draft. All authors: writing—review and editing.

## FUNDING

This work was supported by the Swiss National Science Foundation (Grant Number 407240_167079) in the context of the Swiss National Research Programme 72 Antimicrobial Resistance (www.nfp72.ch).

## REFERENCES

1. WHO. *Global Action Plan on Antimicrobial Resistance*. (2015). Available online at: http://www.who.int/antimicrobial-resistance/publications/global-action-plan/en/ (accessed November 17, 2017).
2. Dyar OJ, Huttner B, Schouten J, Pulcini C, ESGAP (ESCMID Study Group for Antimicrobial stewardshiP). What is antimicrobial stewardship? *Clin Microbiol Infect*. (2017) 23:793–8. doi: 10.1016/j.cmi.2017.08.026
3. Barlam TF, Cosgrove SE, Abbo LM, MacDougall C, Schuetz AN, Septimus EJ, et al. Implementing an antibiotic stewardship program: guidelines by the infectious diseases society of america and the society for healthcare epidemiology of America. *Clin Infect Dis*. (2016) 62:e51–77. doi: 10.1093/cid/ciw118
4. Rawson TM, Moore LSP, Hernandez B, Charani E, Castro-Sanchez E, Herrero P, et al. A systematic review of clinical decision support systems for antimicrobial management: are we failing to investigate these interventions appropriately? *Clin Microbiol Infect*. (2017) 23:524–32. doi: 10.1016/j.cmi.2017.02.028
5. Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, et al. Clinical decision support systems for the practice of evidence-based medicine. *J Am Med Inform Assoc*. (2001) 8:527–34. doi: 10.1136/jamia.2001.0080527
6. Versporten A, Zarb P, Caniaux I, Gros M-F, Drapier N, Miller M, et al. Antimicrobial consumption and resistance in adult hospital inpatients in 53 countries: results of an internet-based global point prevalence survey. *Lancet Glob Health*. (2018) 6:e619–29.
7. Davey P, Marwick CA, Scott CL, Charani E, McNeil K, Brown E, et al. Interventions to improve antibiotic prescribing practices for hospital inpatients. *Cochrane Database Syst Rev*. (2017) 2:CD003543. doi: 10.1002/14651858.CD003543.pub4
8. Catho G, De Kraker M, Waldispühl Suter B, Valotti R, Harbarth S, Kaiser L, et al. Study protocol for a multicentre, cluster randomised, superiority trial evaluating the impact of computerised decision support, audit and feedback on antibiotic use: the COMPuterized Antibiotic Stewardship Study (COMPASS). *BMJ Open*. (2018) 8:e022666. doi: 10.1136/bmjopen-2018-022666
9. *Using Computers to Improve Prescription Practices—NFP*. (2017). Available online at: http://www.nfp72.ch/en/projects/module-3-optimised-use-of-antibiotics/using-computers-to-improve-prescription-practices (accessed July 6, 2020).
10. WHO. *WHO Digital Accelerator Kits*. (2020). Available online at: http://www.who.int/reproductivehealth/publications/digital-accelerator-kits/en/ (accessed March 24, 2020).
11. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*. (2005) 330:765. doi: 10.1136/bmj.38398.500764.8F
12. Wright A, Ash JS, Aaron S, Ai A, Hickman T-TT, Wiesen JF, et al. Best practices for preventing malfunctions in rule-based clinical decision support alerts and reminders: results of a Delphi study. *Int J Med Inf*. (2018) 118:78–85. doi: 10.1016/j.ijmedinf.2018.08.001

13. Catho G, Centemero NS, Catho H, Ranzani A, Balmelli C, Landelle C, et al. Factors determining the adherence to antimicrobial guidelines and the adoption of computerised decision support systems by physicians: a qualitative study in three European hospitals. *Int J Med Inf*. (2020) 141:104233. doi: 10.1016/j.ijmedinf.2020.104233
14. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR Med Inform*. (2018) 6:e24. doi: 10.2196/medinform.8912
15. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc*. (2003) 10:523–30. doi: 10.1197/jamia.M1370
16. Wright A, Goldberg H, Hongsermeier T, Middleton B. A description and functional taxonomy of rule-based decision support content at a large integrated delivery network. *J Am Med Inform Assoc*. (2007) 14:489–96. doi: 10.1197/jamia.M2364
17. Khalifa M, Zabani I. Improving utilization of clinical decision support systems by reducing alert fatigue: strategies and recommendations. *Stud Health Technol Inform*. (2016) 226:51–4. doi: 10.3233/978-1-61499-664-4-51
18. Neo JRJ, Niederdeppe J, Vielemeyer O, Lau B, Demetres M, Sadatsafavi H. Evidence-based strategies in using persuasive interventions to optimize antimicrobial use in healthcare: a narrative review. *J Med Syst*. (2020) 44:64. doi: 10.1007/s10916-020-1531-y
19. Horsky J, Schiff GD, Johnston D, Mercincavage L, Bell D, Middleton B. Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions. *J Biomed Inform*. (2012) 45:1202–16. doi: 10.1016/j.jbi.2012.09.002
20. Buehler R, Griffin D, Peetz J. Chapter one—the planning fallacy: cognitive, motivational, and social origins. In: Zanna MP, Olson JM, editors. *Advances in Experimental Social Psychology*. Academic Press (2010). p. 1–62. Available online at: http://www.sciencedirect.com/science/article/pii/S0065260110430014 (accessed June 26, 2020).
21. Mozaffar H, Cresswell KM, Lee L, Williams R, Sheikh A, On behalf of the NIHR ePrescribing Programme Team. Taxonomy of delays in the implementation of hospital computerized physician order entry and clinical decision support systems for prescribing: a longitudinal qualitative study. *BMC Med Inform Decis Mak*. (2016) 16:25. doi: 10.1186/s12911-016-0263-x
22. WHO. *ICD-10 Online Versions*. (2019). Available online at: http://www.who.int/classifications/icd/icdonlineversions/en/ (accessed July 6, 2020).
23. Li AC, Kannry JL, Kushniruk A, Chrimes D, McGinn TG, Edonyabo D, et al. Integrating usability testing and think-aloud protocol analysis with 'near-live' clinical simulations in evaluating clinical decision support. *Int J Med Inf*. (2012) 81:761–72. doi: 10.1016/j.ijmedinf.2012.02.009
24. Richardson S, Feldstein D, McGinn T, Park LS, Khan S, Hess R, et al. Live usability testing of two complex clinical decision support tools: observational study. *JMIR Hum Factors*. (2019) 6:e12471. doi: 10.2196/12471
25. Richardson S, Mishuris R, O'Connell A, Feldstein D, Hess R, Smith P, et al. 'Think aloud' and 'Near live' usability testing of two complex clinical decision support tools. *Int J Med Inf*. (2017) 106:1–8. doi: 10.1016/j.ijmedinf.2017.06.003
26. de Kraker MEA, Abbas M, Huttner B, Harbarth S. Good epidemiological practice: a narrative review of appropriate scientific methods to evaluate the impact of antimicrobial stewardship interventions. *Clin Microbiol Infect*. (2017) 23:819–25. doi: 10.1016/j.cmi.2017.05.019
27. Teixeira M, Cook DA, Heale BSE, Del Fiol G. Optimization of infobutton design and implementation: a systematic review. *J Biomed Inform*. (2017) 74:10–9. doi: 10.1016/j.jbi.2017.08.010

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2021 Catho, Centemero, Waldispühl Suter, Vernaz, Portela, Da Silva, Valotti, Coray, Pagnamenta, Ranzani, Piuz, Elzi, Meyer, Bernasconi, Huttner and the COMPASS Study Group. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 18 February 2021
doi: 10.3389/fcimb.2020.571230

# Performance of Interferon-Gamma Release Assays in the Diagnosis of Nontuberculous Mycobacterial Diseases—A Retrospective Survey From 2011 to 2019

*Chi Yang, Xuejiao Luo, Lin Fan, Wei Sha, Heping Xiao and Haiyan Cui**

*Shanghai Clinical Research Center for Tuberculosis, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China*

OPEN ACCESS

**Edited by:**
*Adrian Egli, University Hospital of Basel, Switzerland*

**Reviewed by:**
*Meghan Starolis, Quest Diagnostics, United States*
*Riti Sharan, Texas Biomedical Research Institute, United States*

***Correspondence:**
*Haiyan Cui*
*cuihaiyan@tongji.edu.cn*

**Specialty section:**
*This article was submitted to Clinical Microbiology, a section of the journal Frontiers in Cellular and Infection Microbiology*

**Received:** *10 June 2020*
**Accepted:** *28 December 2020*
**Published:** *18 February 2021*

**Citation:**
*Yang C, Luo X, Fan L, Sha W, Xiao H and Cui H (2021) Performance of Interferon-Gamma Release Assays in the Diagnosis of Nontuberculous Mycobacterial Diseases—A Retrospective Survey From 2011 to 2019. Front. Cell. Infect. Microbiol. 10:571230. doi: 10.3389/fcimb.2020.571230*

There is an urgent need for precise diagnosis to distinguish nontuberculous mycobacterial (NTM) diseases from pulmonary tuberculosis (PTB) and other respiratory diseases. The aim of this study is to evaluate the diagnostic performance of Interferon-gamma (IFN-γ) release assays (IGRAs), including antigen-specific peripheral blood-based quantitative T cell assay (T-SPOT.TB) and QuantiFERON-TB-Gold-Test (QFT-G), in differentiating NTM infections ($N$ = 1,407) from culture-confirmed PTB ($N$ = 1,828) and other respiratory diseases ($N$ = 2,652). At specie level, 2.56%, 10.73%, and 16.49% of NTM-infected patients were infected by *Mycobacterium kansasii*, *M. abscessus*, and with *M. avmm-intracellulare* complex (MAC), respectively. Valid analyses of T-SPOT.TB (ESAT-6, CFP-10) and QFT-G were available for 37.03% and 85.79% in NTM-infected patients, including zero and 100% (36/36) of *M. kansasii* infection, 21.85% (33/151) and 92.05% (139/151) of *M. abscessus* infection, and 17.67% (41/232) and 91.24% (211/232) of MAC infection. Based on means comparisons and further ROC analysis, T-SPOT.TB and QFT-G performed moderate accuracy when discriminating NTM from PTB at modified cut-off values (ESAT-6 < 4 SFCs, CFP-10 < 3 SFCs, and QFT-G < 0.667 IU/ml), with corresponding AUC values of 0.7560, 0.7699, and 0.856. At species level of NTM, QFT-G effectively distinguished between MAC (AUC=0.8778), *M. kansasii* (AUC=0.8834) or *M. abscessus* (AUC=0.8783) than T-SPOT.TB. No significant differences in discriminatory power of these three IGRA tools were observed when differentiating NTM and Controls. Our results demonstrated that T-SPOT.TB and QFT-G were both efficient methods for differentiating NTM disease from PTB, and QFT-G possessed sufficient discriminatory power to distinguish infections by different NTM species.

**Keywords: NTM disease, diagnose performance, IGRAs, QFT-G, T-SPOT.TB.**

## INTRODUCTION

*Mycobacteria* are a group of extremely diverse and ubiquitous microorganisms and inhabit nearly every environmental niches (von Reyn et al., 1993; Falkinham, 2002; Johansen et al., 2020), consisting of two major categories: tuberculosis (TB) - causing mycobacteria (MTB) and non-tuberculous mycobacteria (NTM) (Runyon, 1959; Wolinsky, 1992). Partial NTM (*M. avium Complex (MAC), M. kansasii, M. abscessus, M. chelonae, M. fortuitum, M. genavense, M. gordonae, M. haemophilum, M. immunogenum, M. malmoense, M. marinum, M. mucogenicum, M. nonchromogenicum, M. scrofulaceum, M. simiae, M. smegmatis, M. szulgai, M. terrae complex, M. ulcerans, M. xenopi*) are opportunistic pathogens to humans and are the cause of most common lung diseases in clinical with rapidly increasing prevalence worldwide, especially in immuno-compromised patients (Marras et al., 2007; Billinger et al., 2009; Prevots et al., 2010; Thomson et al., 2010; Winthrop et al., 2010; Tsai et al., 2011). The microscopic examination of sputum for acid-fast bacilli (AFB) is a diagnostic standard of pulmonary tuberculosis (PTB). However, the AFB smear-positive are also present in NTM infection. The recovery rate of NTM in AFB positive patients was already considerably high with geographical variation, for instance, 48.5% in the United States (Wright et al., 1998), 43.2% in Australia (Anargyros et al., 1990), 21.1% in Spain (Coll et al., 2003) and 9.1% in Korea (Jeon et al., 2005; Glassroth, 2008; Ryoo et al., 2008). Thus, early clinical identification of NTM infection and PTB would be helpful in patients with AFB smear-positive sputum, as well as for NTM infection and other respiratory diseases with AFB smear-negative sputum (Griffith et al., 2007). However, due to similar clinical symptoms of these lung disease, traditional diagnostic methods, including tuberculin skin test (TST or Mantoux) and chest-X-ray (CXR) are considered unreliable in the diagnosis of MTB. Several molecular techniques (PCR restriction analysis, Anyplex MTB/NTM detection assay, GenoType Mycobacteria Direct test) had been developed for early NTM detection and already been commercially available. However, these tools were regarded to be costly, less sensitive than conventional acid-fast bacilli (AFB) and therefore not recommended in routine clinical practice by British Thoracic Society guidelines at the present time (Haworth et al., 2017). Therefore, there is an urgent need of an early, fast diagnostic technology to distinguish NTM infection from PTB, and from other respiratory diseases (Huebner et al., 1993; Kim et al., 2014).

Interferon-gamma (IFN-γ) release assays (IGRAs), including T-SPOT.TB and QFT-G, display a higher sensitivity compared to the TST for specific detection of latent TB, pulmonary TB or extrapulmonary TB, based on the T-cell mediated IFN-γ release induced by specific *M. tuberculosis* antigens, including ESAT-6, CFP-10 and TB7.7. These specific peptide antigens are usually located in the region of difference (RD1) of MTB genome, and RD1 usually exists in various species of mycobacteria belonging to the *M. tuberculosis* complex (*M. tuberculosis*, *M. bovis*, *M. africanum*, *M. canettii*, *M. caprae*, *M. orygis*, *M. microti*, *M. pinnipedii*, and *M. mungi*) (van Ingen et al., 2012), while only a few species of NTM (*M. kansasii*, *M. gastri*, *M. marinum*, *M. szulgai*, and *M. riyadhense*) share the similar RD1 areas (Harboe et al., 1996; Mahairas et al., 1996; van Ingen et al., 2009). Therefore, IGRAs present high sensitivity for discriminating NTM and MTB.

The aim of this study is to evaluate the efficiency of three different IGRAs (ESAT-6, CFP-10 and QFT-G) for diagnosing NTM infection from PTB and other respiratory diseases.

## MATERIAL AND METHODS

### Patient Population and Ethics Statement

This retrospective study collected clinical data from the Shanghai Pulmonary Hospital, Tongji University School of Medicine (Shanghai, P.R. China) between October 2011 to July 2019. In total, 1,407 consecutive patients diagnosed with NTM infection by culture for mycobacteria were enrolled. 1,828 patients with culture-confirmed pulmonary tuberculosis (PTB) and 2,652 patients with respiratory diseases (pneumonia, pulmonary malignancy, bronchiectasis *etc.*) excluding those infected with NTM or PTB, were enrolled as controls. These respiratory diseases cases without positive results from NTM and mycobacterial culture were mainly diagnosed by etiology, clinical symptoms, imaging findings or pathological examination. The Institutional Review Board of Shanghai Pulmonary Hospital affiliated with Tongji University approved the study and waived the need for informed consent since no patients were at risk. All clinical records were anonymized and de-identified prior to analysis.

### Classification and Diagnosis

NTM diseases were diagnosed with modified guidelines of the American Thoracic Society (ATS) and the Infectious Disease Society of America (IDSA) 2007 criteria (Griffith et al., 2007; Andrejak et al., 2010). Patients with positive culture for NTM from extra pulmonary sites (skin, lymph nodes *etc.*) were also included in line with Freeman et al. (2007). The NTM-infected patients with a previous history of TB disease or MTB isolations from clinical specimens were excluded. Patients were excluded due to discordant IGRAs results, or results not within a 6 month period before or after the positive NTM culture. PTB was diagnosed by sputum culture according to the World Health Organization guidelines (WHO, 2010).

1,407 NTM-infected patients were included including 36 identified as *M. kansasii* infection, 151 as *M. abscessus* infection and 232 as MAC infection. There were 2,652 control patients without NTM and PTB, including 941 (35%) with pneumonia, 599 (23%) with a pulmonary malignancy and 358 (13%) with bronchiectasis. All the participants ($n$ = 198) had negative results on serological tests for human immunodeficiency virus (HIV). The demographic and clinical characteristics of all participants are shown in **Table 1**.

### Laboratory Tests and Examination

All bacterial cultures were assessed using the BD BACTEC™ MGIT™ automated mycobacterial detection system (Becton,

**TABLE 1** | Demographic and clinical characteristics of patients with participants.

| index | *N*TM | Three major *N*TMs | | | PTB | Con*t*rols |
|---|---|---|---|---|---|---|
| | ***N* = 1,407** | ***M.kansasii N* =36** | ***M. abscessus N* =151** | **MAC *N* =232** | ***N* = 1,828** | ***N* = 2,652** |
| Age, years | 60 ± 15 | 50 ± 14 | 59 ± 13 | 60 ± 14 | 43 ± 18 | 56 ± 16 |
| Sex, male | 610(43) | 25(69) | 51(34) | 91(39) | 1,230(67) | 1,593(60) |
| Concomitant diseases (%) | | | | | | |
| Diabetes mellitus | 79(6) | 0(0) | 3(2) | 11(5) | 242(13) | 258(10) |
| Malignancy | 55(4) | 0(0) | 4(3) | 6(3) | 66(4) | 655(25) |
| Rheumatic disease | 39(3) | 1(3) | 4(3) | 8(3) | 19(1) | 43(2) |
| Coronary heart disease | 29(2) | 0(0) | 0(0) | 4(2) | 22(1) | 82(3) |
| Hypertension | 154(11) | 2(6) | 13(9) | 28(1) | 141(8) | 496(19) |

*Data presented as mean ± SD or n of patients (%). N represented for enrolled patient number.*

USA) and all mycobacterial cultures were evaluated using the BD BACTEC™ MGIT™ automated mycobacterial detection system (Becton, Dickinson and Company, Franklin Lakes, NJ, USA). Subsequently, partial species (MAC, *M. kansasii M. abscessus M. gastri*, *M. marinum*, M. szulgai *etc.*) of NTM were identified by 16S rRNA gene sequencing as described previously (Hall et al., 2003). The T-SPOT®.TB assays were conducted following manufacturer's instructions (Oxford, UK). Briefly, all blood samples were collected immediately prior to the tests in order to reduce potential interferences. Peripheral blood mononuclear cells (PBMCs) were isolated from a whole blood sample using Ficoll-Hypaque gradient centrifugation at 400 × g for 30 min at 20°C. Then, the PBMCs were incubated with antigens to stimulate INF-γ secretion by the T cells, and seeded on precoated IFN-γ ELISpot plates followed by incubation with a medium without an antigen (negative control), or a medium containing peptide antigens from ESAT-6 (panel A) or peptide antigens from CFP-10 (panel B), or a medium containing phytohemagglutinin (positive control) in a 5% $CO_2$ atmosphere at 37°C for 20 h (Wang et al., 2010; Wang et al., 2012). The spot-forming cells were counted by an ELISPOT plate reader (AID-GmbH, Straßberg, Germany). Quantitative results for the T-SPOT.TB test are interpreted by subtracting the spot count in the negative controls well from the spot count in each of the Panels, and this number must be at least two-times greater than the spot-forming cells (SFCs) number from the negative wells (Bouwman et al., 2012). All tests were performed before anti-TB medication. The QuantiFERON-TB-Gold-Test (Cellestis Ltd., Carnegie, Victoria, Australia) was also performed following the manufacturer's recommendations. Briefly, aliquots of heparinized whole blood are incubated with the test antigens (ESAT-6, CFP-10, and TB 7.7 proteins) for 16–24 h; phytohemaglutinin is performed as the positive assay control, and saline as the negative control (nil tube). After incubation, the concentration of IFN-g in the plasma would be read by ELISA and the quantitative result of the test was reported as the IFN-γ level in the sample tube minus the baseline level (nil tube) (Mazurek et al., 2005).

## Statistical Analyses

All data were analyzed by using MedCalc® version 9.0.1.1 (MedCalc, Belgium). The results by different tests were compared using χ2-test to assess the potential to discriminate each other. Then, the ROC curves were calculated between the groups with statistically significant difference. Areas under the ROC curve (AUC) are evaluated to assess the discriminatory powers of IGRA test. Generally, the AUC values are positive correlation to reliability and discrimination, in which higher than 0.9 indicates high accuracy, 0.7–0.9 indicates moderate accuracy, 0.5–0.7 indicates low accuracy, and less than 0.5 indicates no discrimination (Oh et al., 1993; Fischer et al., 2003). An AUC value greater than 0.7 on the validation can be considered as acceptable models for differentiation. The corresponding index, including optimal cut-off values, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) are calculated.

# RESULTS

In total, valid results of T-SPOT.TB were available for 37.03% (521/1,407) in NTM-infected patients, 30.03% (549/1,828) in PTB patients and 48.57% (1,288/2,652) in controls. For QFT-G, valid results were available for 73.13% (1,029/1,407) in NTM-infected patients, 69.97% (1,279/1,828) in PTB patients and 51.43% (1,364/2,652) in controls. At the species level of NTM, the valid results of T-SPOT.TB were available for 21.85% (33/151) in patients infected with *M. abscessus* and 17.67% (41/232) with MAC. For QFT-G, valid results were available for 100% (36/36) of patients infected with *M. kansasii*, 92.05% (139/151) with *M. abscessus* and 90.95% (211/232) with MAC (**Table S1**). Quantitative results of different IGRAs were compared (**Table S2**). The *P* values indicated both T-SPOT.TB and QFT-G were effective when discriminating NTM form PTB ($P < 0.001$), while QFT-G showed lower performance ($P = 0.1853$) than T-SPOT.TB ($P = 0.0000$ for ESAT-6 and $P = 0.0017$ for CFP-10) when discriminating NTM from controls.

To further evaluate the diagnostic performance of IGRAs when discriminating NTM from PTB, the ROC analyses with statistical significances were conducted. As shown in **Table 2**, ESAT-6 (AUC: 0.7560), CFP-10 (AUC: 0.7699) and QFT-G (AUC: 0.8560) had moderate accuracy with AUC > 0.7 (**Figure S1**). Notably, based on sensitivity, specificity, PPV, NPV, and AUC, QFT-G showed better diagnostic performance than ESAT-6 and CFP-10. When discriminating NTM from Controls, ESAT-6 and CFP-10 displayed low accuracy with AUC < 0.7 (**Figure S2**, **Table 3**). At the species level of NTM, ESAT-6 and CFP-10 both had moderate accuracy with AUC > 0.7 when discriminating MAC or *M abscessus* from PTB. QFT-G also performed moderate

**TABLE 2 |** Diagnostic performance of interferon-gamma release assays (IGRA) tools to discriminate nontuberculous mycobacterial (NTM) from pulmonary tuberculosis (PTB).

| Index | QFT-G | ESAT-6 | CFP-10 |
|---|---|---|---|
| **Cut-off value to distinguish NTM** | <0.667 IU/ml | <4 SFCs | <3 SFCs |
| **Sensitivity (%)** | 0.867 | 0.7942 | 0.7596 |
| **Specificity (%)** | 0.7483 | 0.6315 | 0.6756 |
| **PPV (%)** | 0.7349 | 0.6716 | 0.6897 |
| **NPV (%)** | 0.8750 | 0.7638 | 0.7475 |
| **AUC** | 0.8560 | 0.7560 | 0.7699 |
| **95%CI** | 0.840–0.8719 | 0.7268–0.7852 | 0.7415–0.7983 |
| ***P*-value** | 0.0000 | 0.0000 | 0.0000 |

**TABLE 3 |** Diagnostic performance of T-SPOT.TB to discriminate nontuberculous mycobacterial (NTM) from controls.

| Index | ESAT-6 | CFP-10 |
|---|---|---|
| **Cut-off value** | >1 SFCs | >2 SFCs |
| **Sensitivity (%)** | 0.4664 | 0.3244 |
| **Specificity (%)** | 0.6421 | 0.7539 |
| **PPV (%)** | 0.3452 | 0.3477 |
| **NPV (%)** | 0.7484 | 0.7339 |
| **AUC (95% CI)** | 0.5537 | 0.5458 |
| **95% CI** | 0.5236–0.5838 | 0.5160–0.5756 |
| ***P*-value** | P=0.0003 | P=0.002 |

accuracy with AUC > 0.7 when discriminating MAC, *M. kansasii* or *M. abscessus* from PTB, However, QFT-G established better performance either ESAT-6 or CFP-10 when discriminating MAC or *M. abscessus* from PTB (**Figure S3**, **Table 4**).

## DISCUSSION

Previous reports about the diagnostic performance of IGRAs in distinguishing NTM infections remain limited and their conclusions were very heterogeneous. Hermansen reported the QFT-G positivity rate was 8% (4/53) in definite NTM disease and 31% (15/49) in possible disease with colonization, while the overall rate of positive QFT-G in pulmonary NTM disease defined patients was 18% (81/462) by their systematic review (Hermansen et al., 2014). Augustynowicz-Kopeć reported that the positive IGRAs result was 8% in NTM definite patients (3/39) (Augustynowicz-Kopeć et al., 2019). Wang stated that the positivity rate of T-SPOT.TB was 53.4% (31/58) among the probable and definite NTM groups, 53.5% (15/28) for probable cases and 53.3% (16/30) for definite cases (Wang et al., 2016). Siegel reported the lowest positivity rate (2.0%, 1/51) for both QFT-G assay and a new-generation QuantiFERON-TB Gold Plus (QFT-Plus) assay in patients with MAC or *M. abscessus* (Siegel et al., 2018). In our article, with the traditional cut-off, positivity rate for QFT-G assay and T-SPOT.TB would be respectively 30.8% (317/1029) and 34.7% (181/522). The heterogeneity of these studies might be ascribed into several reasons as follow.

Firstly, the case number with valid IGRA results (all less than 100 from previous reports) limited the accuracy of analyses and comparison. Secondly, the lack of completed control group (PTB and other respiratory diseases) for comparisons may limit the estimation the discriminating power. Thirdly, these researches evaluated the discriminating power for NTM by calculating the positivity rate of IGRAs with cut-off values designed to differentiate TB, while distinguishing NTM, a modified cut-off

**TABLE 4 |** Diagnostic performance of interferon-gamma release assay (IGRA) tools to discriminate species of nontuberculous mycobacterial (NTM) from pulmonary tuberculosis (PTB).

| | Index | QFT-G | ESAT-6 | CFP-10 |
|---|---|---|---|---|
| **MAC from PTB** | Cut-off value | <0.6670 IU/ml | <2 SFCs | <2 SFCs |
| | Sensitivity (%) | 0.8671 | 0.8525 | 0.7996 |
| | Specificity (%) | 0.8389 | 0.6829 | 0.7805 |
| | PPV (%) | 0.4703 | 0.1672 | 0.2139 |
| | NPV (%) | 0.9745 | 0.9841 | 0.9812 |
| | AUC (95%CI) | 0.8778 | 0.8226 | 0.8183 |
| | 95% CI | 0.8479–0.9077 | 0.7583–0.8868 | 0.7536–0.8830 |
| | *P* value | 0.0000 | 0.0000 | 0.0000 |
| ***M. kansasii* from PTB** | Cut-off value | <0.8180 IU/ml | – | – |
| | Sensitivity (%) | 0.8444 | – | – |
| | Specificity (%) | 0.8611 | – | – |
| | PPV (%) | 0.1461 | – | – |
| | NPV (%) | 0.9949 | – | – |
| | AUC (95%CI) | 0.8834 | – | – |
| | 95% CI | 0.8315–0.9354 | – | – |
| | *P* value | 0.0000 | – | – |
| ***M. abscessus* from PTB** | Cut-off value | <0.3565 IU/ml | <4 SFCs | <1 SFCs |
| | Sensitivity (%) | 0.8999 | 0.7942 | 0.8452 |
| | Specificity (%) | 0.8058 | 0.7576 | 0.7879 |
| | PPV (%) | 0.3349 | 0.1645 | 0.1932 |
| | NPV (%) | 0.9867 | 0.9839 | 0.9883 |
| | AUC (95%CI) | 0.8783 | 0.8199 | 0.8383 |
| | 95% CI | 0.8417–0.9149 | 0.7410–0.8987 | 0.7690–0.9076 |
| | *P* value | 0.0000 | 0.0000 | 0.0000 |

*–, not available.*

should be optimized by ROC technique in the target population (Greiner et al., 2000).

In our research, we finally obtained 1029 valid results of QFT-G and 521 valid results of T-SPOT.TB among almost 80,000 patients enrolled from 2011 to 2019 with potential NTM disease, which was 10 times larger than the research by Andrejak (Andrejak et al., 2010) and 10 times larger than the research by Wang in China (Wang et al., 2016). Based on a large consecutive data set, we evaluated the discriminatory power of IGRAs between NTM and PTB with the area under the ROC curve (AUC), which is regarded as a global summary statistic of diagnostic accuracy (Greiner et al., 2000). Our data suggested a moderate accuracy with AUC > 0.7 for IGRAs to differentiate these diseases. Therefore, we recommended the application of IGRA tools (T-SPOT. TB and QFT-G) to distinguish NTM in AFB smear-positive patients who were composed of only NTM and PTB patients, and QFT-G may be preferred due to its higher AUC value.

Furthermore, our data showed that no significant difference of discrimination power was identified between NTM and Control patients by IGRAs. Since smear-negative patient group was composed of NTM, PTB, latent tuberculosis infection (LTBI) and other respiratory diseases, and the IGRA tools may only differentiate PTB and LTBI from this group, the remaining NTM and other respiratory diseases could not been further differentiated by this method. This result indicated the limitation of IGRAs in distinguishing NTM from AFB smear-negative patients Therefore, our study recommended the application of IGRAs to diagnose NTM in AFB smear-positive patients.

A number of methods detecting NTM from respiratory samples had been applicable, including culture and several molecular techniques. Culture usually provides the most reliable evidence for diagnosis, however, it is time-consuming and may provide negative results even in AFB smear-positive patients, making early diagnosis difficult (Haworth et al., 2017). The molecular techniques (PCR restriction analysis, Anyplex MTB/NTM detection assay, GenoType Mycobacteria Direct test) could be faster and highly specific but still share the limitation to be occasionally ineligible (Haworth et al., 2017). The PCR restriction analysis was reported to successfully amplify mycobacterial DNA in only 60% (72/121) of NTM patients with AFB smear-positive sputum Kim et al., 2008). Franco-Alvarez de Luna et al. reported that the GenoType Mycobacteria Direct test, which is capable of detecting TB and four atypical mycobacterium species, detected 92% (93/101) of tuberculosis patients and 22% (6/27) of nontuberculosis patients in culture positive samples (Franco-Alvarez De Luna et al., 2006). Perry et al. showed that the rate of negative results from a molecular tool (Anyplex MTB/NTM detection assay) was 11% (10/91) in smear positive patients (Perry et al., 2014). Kim et al. also reported two real time PCR assays (Anyplex plus MTB/NTM Detection kit and Genedia MTB/NTM Detection kit) had respectively 82%(14/17) and 76%(13/17) positive rate in NTM culture positive patients (Kim et al., 2020). Shin et al. reported the positive rate of Genedia MTB/NTM Detection kit was only 23%(16/69) in NTM culture positive patients (Shin et al., 2020). In our research, IGRA tools (QFT-G, ESAT-6 and CFP-10) could detect 87%, 79% and 76% NTM in smear-positive patients, and still could play an alternative or complementary role in discrimination of NTM, especially while negative test results occur with other tools. A suggested algorithm for the investigation of smear positive individuals is shown in **Figure 1**.

Previous studies usually adopted the recommended cut off value of IGRAs which were designed for differentiating PTB from Controls, and some researchers already revealed that a revised cut-off value may increase the sensitivity to detect NTM disease. Kobashi et al. reported when the cut-off value of positive response for QFT-G changed from 0.35 to 0.20 IU/ml, the sensitivity to detect NTM disease (*M. kansasii* disease) increased from 52% to 82%, while specificity decreased from 93% to 91% (Kobashi et al., 2009). Similarly, our survey showed that modified cut-off values for discriminating NTM (ESAT-6 < 4 SFCs; CFP-10 < 3 SFCs and QFT-G < 0.667 IU/mL) could also improve their summary accuracies. For QFT-G, its sensitivity would decrease from 90% to 87%, while specificity would increase from 69% to 75%. For T-SPOT, its sensitivities for ESAT-6 and CFP-10 would grow from 70% and 66% to 79% and 76%, respectively; correspondingly their specificities would decrease from 71% and 77% to 63% and 68%.

Meanwhile, some NTM species (*M. kansasii*, *M. marinum* and *M. szulgai*) sharing RD1 with *M. tuberculosis* showed similar positive results with IGRAs as *M. tuberculosis* (Hermansen et al., 2014; Chen et al., 2018; Augustynowicz-Kopeć et al., 2019), and the feasibility of differentiating these species from PTB was doubtful. Our analysis suggested there was no statistically significant difference of IGRAs performances between any species of NTM. Additionally, the ROC curve in this study showed that QFT-G displayed a moderate accuracy (AUC = 0.8834) when distinguishing some RD1-posesssing NTM species (at least *M. kansasii*), which was inconsistent with other researchers and might be caused by a revised cut-off (< 0.8180 IU/mL) to amplify the difference between these RD1-posesssing NTM species patients and PTB patients.

Last, we need to emphasize the limitations of our work. First, the patients were collected from a single center in a high TB incidence country, which may limit its generalizability in low TB incidence country. More intercontinental surveys are needed to expand its universality. Second, as a retrospective review, bacterial species identification could not be performed in all the patients. Therefore, the analysis of discriminatory power was unavailable for species with scarce isolations, including *M. gastri*, *M. marinum*, *M. szulgai* and so on. Third, due to the retrospective nature of this article, the BCG vaccination statuses of our patients are unavailable. BCG vaccination was regarded to be associated with durable IFN-g responses and its impact on the performances of IGRA tools should not be neglected.

To our best knowledge, this study is one of the largest assessing IGRAs with valid results in discriminating NTM infections from both AFB smear positive (PTB) and AFB smear negative patients. T-SPOT.TB and QFT-G were performed in patients with NTM infection, PTB and other respiratory diseases. Our results revealed that, with modified cut-off values, these IGRAs possessed the potential in

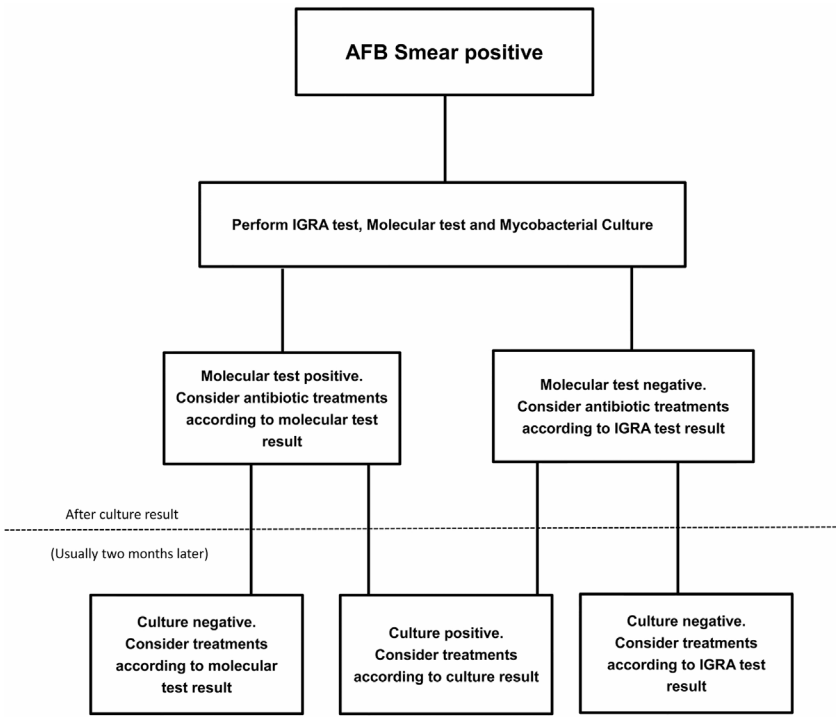



**FIGURE 1** | A suggested algorithm for the nontuberculous mycobacterial (NTM) investigation of smear positive individuals.

differentiating NTM disease from PTB disease in AFB smear-positive patients. Furthermore, for some species of NTM (MAC, *M. abscessus*, even RD1 possessing mycobacteria), the T-SPOT.TB or QFT-G had moderate discriminatory power. However, since many respiratory diseases (pneumonia et al.) in AFB smear-negative patients share similar IGRAs results with NTM, the discrimination power of IGRA tools for NTM in AFB smear-negative patients may be limited in diagnosis. In conclusion, our study provided new insights into the diagnostic performance of IGRAs in differentiation of NTM infection and PTB, and provided more guidance to promote the diagnostic accuracy of PTB and NTM infection in the clinic.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Institutional Review Board of Shanghai Pulmonary Hospital affiliated with Tongji University. Written informed consent from the participants was not required to participate in this study in accordance with the national legislation and the institutional requirements

## AUTHOR CONTRIBUTIONS

HC, CY, HX, and WS contributed to study design. CY, XL, LF, HX, and WX performed the experiments. CY and XL performed data analysis. CY and HC wrote the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This work is supported by Shanghai top priority of clinical medical center and key discipline construction plan (2017ZZ02003), Shanghai science and technology commission project (20Y11901500) and Shanghai clinical research center for infectious disease (tuberculosis, 19MC1910800).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2020.571230/full#supplementary-material

## REFERENCES


Anargyros, P., Astill, D. S., and Lim, I. S. (1990). Comparison of improved BACTEC and Lowenstein-Jensen media for culture of mycobacteria from clinical specimens. *J. Clin. Microbiol.* 28, 1288–1291. doi: 10.1128/JCM.28.6.1288-1291.1990

Andrejak, C., Thomsen, V. O., Johansen, I. S., Riis, A., Benfield, T. L., Duhaut, P., et al. (2010). Nontuberculous pulmonary mycobacteriosis in Denmark: incidence and prognostic factors. *Am. J. Respir. Crit. Care Med.* 181, 514–521. doi: 10.1164/rccm.200905-0778OC

Augustynowicz-Kopeć, E., Siemion-Szczesniak, I., Zabost, A., Wyrostkiewicz, D., Filipczak, D., Oniszh, K., et al. (2019). Interferon gamma release assays in patients with respiratory isolates of non-tuberculous mycobacteria - a preliminary study. *Pol. J. Microbiol.* 68, 15–19. doi: 10.21307/pjm-2019-002

Billinger, M. E., Olivier, K. N., Viboud, C., de Oca, R. M., Steiner, C., Holland, S. M., et al. (2009). Nontuberculous mycobacteria-associated lung disease in hospitalized persons, United States 1998-2005. *Emerg. Infect. Dis.* 15, 1562–1569. doi: 10.3201/eid1510.090196

Bouwman, J. J., Thijsen, S. F., and Bossink, A. W. (2012). Improving the timeframe between blood collection and interferon gamma release assay using T-Cell Xtend(R). *J. Infect.* 64, 197–203. doi: 10.1016/j.jinf.2011.10.017

Chen, Y., Jiang, H., Zhang, W., Chen, Z., Mei, Y., Chen, H., et al. (2018). Diagnostic value of T-SPOT.TB test in cutaneous mycobacterial infections. *Acta Derm. Venereol.* 98, 989–990. doi: 10.2340/00015555-3011

Coll, P., Garrigo, M., Moreno, C., and Marti, N. (2003). Routine use of gen-probe amplified *Mycobacterium tuberculosis* direct (MTD) test for detection of *Mycobacterium tuberculosis* with smear-positive and smear-negative specimens. *Int. J. Tuberc. Lung. Dis.* 7, 886–891.

Falkinham, J.O. III. (2002). Nontuberculous mycobacteria in the environment. *Clin. Chest. Med.* 23, 529–551. doi: 10.1016/S0272-5231(02)00014-X

Fischer, J. E., Bachmann, L. M., and Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive. Care Med.* 29, 1043–1051. doi: 10.1007/s00134-003-1761-8

Franco-Alvarez De Luna, F., Ruiz, P., Gutierrez, J., and Casal, M. (2006). Evaluation of the GenoType Mycobacteria Direct assay for detection of Mycobacterium tuberculosis complex and four atypical mycobacterial species in clinical samples. *J. Clin. Microbiol.* 44, 3025–3027. doi: 10.1128/JCM.00068-06

Freeman, J., Morris, A., Blackmore, T., Hammer, D., Munroe, S., and McKnight, L. (2007). Incidence of nontuberculous mycobacterial disease in New Zealand. *N. Z. Med.* 120, U2580.

Glassroth, J. (2008). Pulmonary disease due to nontuberculous mycobacteria. *Chest* 133, 243–251. doi: 10.1378/chest.07-0358

Greiner, M., Pfeiffer, D., and Smith, R. D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev. Vet. Med.* 45, 23–41. doi: 10.1016/S0167-5877(00)00115-X

Griffith, D. E., Aksamit, T., Brown-Elliott, B. A., Catanzaro, A., Daley, C., Gordin, F., et al. (2007). An official ATS/IDSA statement: diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases. *Am. J. Respir. Crit. Care Med.* 175, 367–416. doi: 10.1164/rccm.200604-571ST

Hall, L., Doerr, K. A., Wohlfiel, S. L., and Roberts, G. D. (2003). Evaluation of the MicroSeq system for identification of mycobacteria by 16S ribosomal DNA sequencing and its integration into a routine clinical mycobacteriology laboratory. *J. Clin. Microbiol.* 41, 1447–1453. doi: 10.1128/JCM.41.4.1447-1453.2003

Harboe, M., Oettinger, T., Wiker, H. G., Rosenkrands, I., and Andersen, P. (1996). Evidence for occurrence of the ESAT-6 protein in *Mycobacterium tuberculosis* and virulent *Mycobacterium bovis* and for its absence in *Mycobacterium bovis* BCG. *Infect. Immun.* 64, 16–22. doi: 10.1128/IAI.64.1.16-22.1996

Haworth, C. S., Banks, J., Capstick, T., Fisher, A. J., Gorsuch, T., Laurenson, I. F., et al. (2017). British Thoracic Society guidelines for the management of Non-tuberculous mycobacterial pulmonary disease (NTM-PD). *Thorax* 72, ii1–ii64. doi: 10.1136/thoraxjnl-2017-210929

Hermansen, T. S., Thomsen, V. O., Lillebaek, T., and Ravn, P. (2014). Non-tuberculous mycobacteria and the performance of interferon gamma release assays in Denmark. *PloS. One* 9, e93986. doi: 10.1371/journal.pone.0093986

Huebner, R. E., Schein, M. F., and Bass, J. B. Jr. (1993). The tuberculin skin test. *Int. J. Tuberc. Lung. Dis.* 17, 968–975. doi: 10.1093/clinids/17.6.968

Jeon, K., Koh, W. J., Kwon, O. J., Suh, G. Y., Chung, M. P., Kim, H., et al. (2005). Recovery rate of NTM from AFB smear-positive sputum specimens at a medical centre in South Korea. *Int. J. Tuberc. Lung. Dis.* 9, 1046–1051.

Johansen, M. D., Herrmann, J.-L., and Kremer, L. (2020). Non-tuberculous mycobacteria and the rise of *Mycobacterium abscessus*. *Nat. Rev. Microbiol.* 18 (7), 392–407. doi: 10.1038/s41579-020-0331-1

Kim, S., Park, E. M., Kwon, O. J., Lee, J. H., Ki, C. S., Lee, N. Y., et al. (2008). Direct application of the PCR restriction analysis method for identifying NTM species in AFB smear-positive respiratory specimens. *Int. J. Tuberc. Lung. Dis.* 12, 1344–1346.

Kim, Y. K., Hahn, S., Uh, Y., Im, D. J., Lim, Y. L., Choi, H. K., et al. (2014). Comparable characteristics of tuberculous and non-tuberculous mycobacterial cavitary lung diseases. *Int. J. Tuberc. Lung. Dis.* 18, 725–729. doi: 10.5588/ijtld.13.0871

Kim, J., Choi, Q., Kim, J. W., Kim, S. Y., Kim, H. J., Park, Y., et al. (2020). Comparison of the Genedia MTB/NTM Detection Kit and Anyplex plus MTB/NTM Detection Kit for detection of Mycobacterium tuberculosis complex and nontuberculous mycobacteria in clinical specimens. *J. Clin. Lab. Anal.* 34, e23021. doi: 10.1002/jcla.23021

Kobashi, Y., Mouri, K., Yagi, S., Obase, Y., Miyashita, N., Okimoto, N., et al. (2009). Clinical evaluation of the QuantiFERON-TB Gold test in patients with non-tuberculous mycobacterial disease. *Int. J. Tuberc. Lung. Dis.* 13, 1422–1426.

Mahairas, G. G., Sabo, P. J., Hickey, M. J., Singh, D. C., and Stover, C. K. (1996). Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent *M. bovis*. *J. Bacteriol.* 178, 1274–1282. doi: 10.1128/JB.178.5.1274-1282.1996

Marras, T. K., Chedore, P., Ying, A. M., and Jamieson, F. (2007). Isolation prevalence of pulmonary non-tuberculous mycobacteria in Ontario 1997 2003. *Thorax* 62, 661–666. doi: 10.1136/thx.2006.070797

Mazurek, G. H., Jereb, J., Lobue, P., Iademarco, M. F., Metchock, B., Vernon, A., et al. (2005). Guidelines for using the QuantiFERON-TB Gold test for detecting Mycobacterium tuberculosis infection, United States. *MMWR. Recomm. Rep.* 54, 49–55. doi: 10.2307/42000946

Oh, T. E., Hutchinson, R., Short, S., Buckley, T., Lin, E., and Leung, D. (1993). Verification of the acute physiology and chronic health evaluation scoring system in a Hong Kong intensive care unit. *Crit. Care Med.* 21, 698–705. doi: 10.1097/00003246-199305000-00013

Perry, M. D., White, P. L., and Ruddy, M. (2014). Potential for use of the Seegene Anyplex MTB/NTM real-time detection assay in a regional reference laboratory. *J. Clin. Microbiol.* 52, 1708–1710. doi: 10.1128/JCM.03585-13

Prevots, D. R., Shaw, P. A., Strickland, D., Jackson, L. A., Raebel, M. A., Blosky, M. A., et al. (2010). Nontuberculous mycobacterial lung disease prevalence at four integrated health care delivery systems. *Am. J. Respir. Crit. Care Med.* 182, 970–976. doi: 10.1164/rccm.201002-0310OC

Runyon, E. H. (1959). Anonymous mycobacteria in pulmonary disease. *Med. Clin. North. Am.* 43, 273–290. doi: 10.1016/S0025-7125(16)34193-1

Ryoo, S. W., Shin, S., Shim, M. S., Park, Y. S., Lew, W. J., Park, S. N., et al. (2008). Spread of nontuberculous mycobacteria from 1993 to 2006 in Koreans. *J. Clin. Lab. Anal.* 22, 415–420. doi: 10.1002/jcla.20278

Shin, S., Yoo, I. Y., Shim, H. J., Kang, O. K., Jhun, B. W., Koh, W. J., et al. (2020). Diagnostic Performance of the GENEDIA MTB/NTM Detection Kit for Detecting Mycobacterium tuberculosis and Nontuberculous Mycobacteria With Sputum Specimens. *Ann. Lab. Med.* 40, 169–173. doi: 10.3343/alm.2020.40.2.169

Siegel, S. A. R., Cavanaugh, M., Ku, J. H., Kawamura, L. M., and Winthrop, K. L. (2018). Specificity of QuantiFERON-TB Plus, a new-generation interferon gamma release assay. *J. Clin. Microbiol.* 56, e00629-18. doi: 10.1128/JCM.00629-18

Thomson, R. M.Centre NTM working group at Queensland TB Control Centre and Queensland Mycobacterial Reference Laboratory (2010). Changing epidemiology of pulmonary nontuberculous mycobacteria infections. *Emerg. Microbes. Infect.* 16, 1576–1583. doi: 10.3201/eid1610.091201

Tsai, C. F., Shiau, M. Y., Chang, Y. H., Wang, Y. L., Huang, T. L., Liaw, Y. C., et al. (2011). Trends of mycobacterial clinical isolates in Taiwan. *Trans. R. Soc Trop. Med. Hyg.* 105, 148–152. doi: 10.1016/j.trstmh.2010.11.005

van Ingen, J., de Zwaan, R., Dekhuijzen, R., Boeree, M., and van Soolingen, D. (2009). Region of difference 1 in nontuberculous *Mycobacterium* species adds a

phylogenetic and taxonomical character. *J. Bacteriol.* 191, 5865–5867. doi: 10.1128/JB.00683-09

van Ingen, J., Rahim, Z., Mulder, A., Boeree, M. J., Simeone, R., Brosch, R., et al. (2012). Characterization of *Mycobacterium orygis* as *M. tuberculosis* complex subspecies. *Emerg. Infect. Dis.* 18, 653–655. doi: 10.3201/eid1804.110888

von Reyn, C. F., Waddell, R. D., Eaton, T., Arbeit, R. D., Maslow, J. N., Barber, T. W., et al. (1993). Isolation of *Mycobacterium avium* complex from water in the United States, Finland, Zaire, and Kenya. *J. Clin. Microbiol.* 31, 3227–3230. doi: 10.1128/JCM.31.12.3227-3230.1993

Wang, S. H., Powell, D. A., Nagaraja, H. N., Morris, J. D., Schlesinger, L. S., and Turner, J. (2010). Evaluation of a modified interferon-gamma release assay for the diagnosis of latent tuberculosis infection in adult and paediatric populations that enables delayed processing. *Scand. J. Infect. Dis.* 42, 845–850. doi: 10.3109/00365548.2010.498021

Wang, S. H., Stew, S. S., Cyktor, J., Carruthers, B., Turner, J., and Restrepo, B. II. (2012). Validation of increased blood storage times with the T-SPOT.TB assay with T-Cell Xtend reagent in individuals with different tuberculosis risk factors. *J. Clin. Microbiol.* 50, 2469–2471. doi: 10.1128/JCM.00674-12

Wang, M. S., Wang, J. L., and Wang, X. F. (2016). The performance of interferon-gamma release assay in nontuberculous mycobacterial diseases: a retrospective study in China. *BMC. Pulm. Med.* 16, 163. doi: 10.1186/s12890-016-0320-3

Winthrop, K. L., McNelley, E., Kendall, B., Marshall-Olson, A., Morris, C., Cassidy, M., et al. (2010). Pulmonary nontuberculous mycobacterial disease prevalence and clinical features: an emerging public health disease. *Am. J. Respir. Crit. Care Med.* 182, 977–982. doi: 10.1164/rccm.201003-0503OC

Wolinsky, E. (1992). Mycobacterial diseases other than tuberculosis. *Clin. Infect. Dis.* 15, 1–10. doi: 10.1093/clinids/15.1.1

World Health Organization (2010). *Treatment of Tuberculosis: Guidelines. 4th edition* (Geneva: WHO Press). Available at: https://www.ncbi.nlm.nih.gov/books/NBK138748/.

Wright, P. W., Wallace, R. J. Jr., Wright, N. W., Brown, B. A., and Griffith, D. E. (1998). Sensitivity of fluorochrome microscopy for detection of *Mycobacterium tuberculosis* versus nontuberculous mycobacteria. *J. Clin. Microbiol.* 36, 1046–1049. doi: 10.1128/JCM.36.4.1046-1049.1998

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2021 Yang, Luo, Fan, Sha, Xiao and Cui. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Digital Health

MINI REVIEW
published: 03 March 2021
doi: 10.3389/fdgth.2021.583468

# Digital Insights Into Nucleotide Metabolism and Antibiotic Treatment Failure

*Allison J. Lopatkin*[1,2,3*†] *and Jason H. Yang*[4,5*†]

[1] Department of Biology, Barnard College, New York, NY, United States, [2] Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY, United States, [3] Data Science Institute, Columbia University, New York, NY, United States, [4] Ruy V. Lourenço Center for Emerging and Re-emerging Pathogens, Rutgers New Jersey Medical School, Newark, NJ, United States, [5] Department of Microbiology, Biochemistry and Molecular Genetics, Rutgers New Jersey Medical School, Newark, NJ, United States

Nucleotide metabolism plays a central role in bacterial physiology, producing the nucleic acids necessary for DNA replication and RNA transcription. Recent studies demonstrate that nucleotide metabolism also proactively contributes to antibiotic-induced lethality in bacterial pathogens and that disruptions to nucleotide metabolism contributes to antibiotic treatment failure in the clinic. As antimicrobial resistance continues to grow unchecked, new approaches are needed to study the molecular mechanisms responsible for antibiotic efficacy. Here we review emerging technologies poised to transform understanding into why antibiotics may fail in the clinic. We discuss how these technologies led to the discovery that nucleotide metabolism regulates antibiotic drug responses and why these are relevant to human infections. We highlight opportunities for how studies into nucleotide metabolism may enhance understanding of antibiotic failure mechanisms.

**Keywords: antibiotic resistance, antibiotic tolerance, antibiotic persistence, nucleotide metabolism, whole genome sequencing, machine learning, metabolic modeling, predictive modeling**

OPEN ACCESS

***Edited by:***
*Belén Rodriguez-Sanchez, Gregorio Marañón Hospital, Spain*

***Reviewed by:***
*Juan Liu, Huazhong University of Science and Technology, China*
*Hao Wang, Shenzhen University General Hospital, China*

***Correspondence:**
*Allison J. Lopatkin*
*alopatki@barnard.edu*
*Jason H. Yang*
*jason.y@rutgers.edu*

*† These authors have contributed equally to this work*

***Specialty section:***
*This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health*

***Received:*** *15 July 2020*
***Accepted:*** *02 February 2021*
***Published:*** *03 March 2021*

***Citation:***
*Lopatkin AJ and Yang JH (2021) Digital Insights Into Nucleotide Metabolism and Antibiotic Treatment Failure. Front. Digit. Health 3:583468. doi: 10.3389/fdgth.2021.583468*

## INTRODUCTION

In the nearly 100 years since the discovery of penicillin, antibiotics have revolutionized medical practice and have become a cornerstone of modern medicine. However, growing rates of antimicrobial resistance pose an urgent and looming threat to public health and economic stability (1). These are compounded by a diminished antimicrobial discovery pipeline (2), creating a critical need to understand mechanisms responsible for antibiotic treatment failures and to discover new effective antimicrobials.

Clinical microbiology traditionally relies on general microbiology and molecular biology laboratory techniques, such as polymerase chain reaction and gene deletion/over-expression, to elucidate molecular mechanisms responsible for clinical phenotypes. However, experimental throughput by these methods limits progress toward understanding mechanisms of antibiotic treatment failure. In recent years several new experimental and digital technologies have emerged with promise to increase clinical microbiology laboratory throughput and enhance clinical management of bacterial infections (3–5). Moreover, advances in prokaryotic systems biology (6, 7) and interpretable machine learning (8) are for the first time accelerating discovery of mechanisms underlying antibiotic efficacy (9, 10).

Here, we review emerging digitalization technologies poised to transform research into mechanisms of antibiotic treatment failure in the clinic. We describe several antibiotic resistance, tolerance and persistence mechanisms discovered from clinical strains. We discuss in detail

the recent discovery that nucleotide metabolism actively participates in antibiotic lethality and the clinical relevance of these findings (11). We propose new opportunities for digitalization technologies to advance clinical practice and to open frontiers for basic research into nucleotide metabolism and antibiotic efficacy.

## DIGITALIZATION IN CLINICAL AND RESEARCH SETTINGS

The most important goal in clinical microbiology is to identify an infectious pathogen and determine its drug susceptibility profile (12). Traditionally, clinical microbiology laboratories rely on culture-based methods for pathogen identification and susceptibility testing. These approaches require the successful isolation and culture of pathogen cells from a clinical sample, followed by *in vitro* screening with standardized antibiotics.

*In vitro* studies in research settings have enabled the discovery of antibiotic resistance mechanisms. For example, following the initial detection of clinical tetracycline resistance, several microbiology studies identified decreased drug transport as the mechanism responsible for reduced efficacy (13, 14). Subsequent studies identified multi-drug resistant efflux pumps in multiple pathogenic species (e.g., AcrB in *Escherichia coli* and MexB in *Pseudomonas aeruginosa*) (15). As with their clinical counterparts, these fundamental studies rely on culture-based growth and targeted sequencing; however, such experimental technologies are resource- and labor-intensive and do not scale well with the plethora of pathogen variants, drug mechanisms, and resistance strategies found in the clinic.

In recent years, advances in laboratory evolution, high-throughput sequencing, and computational biology have greatly expanded the scope of addressable questions in microbiology and the study of antibiotic resistance (16). For instance, adaptive laboratory evolution can simulate natural selection pressures (17), allowing researchers to study the emergence of novel antibiotic treatment phenotypes (18), as well as their relationship to environmental conditions (19). In many cases, these granular experimental techniques invite complementary computational modeling activities, from mechanistically simulating drug-target binding to predicting complex ecological dynamics, yielding deeper insights into clinical resistance phenomena.

Concurrently, whole-genome sequencing has transformed the study of antibiotic resistance, enabling the identification of all possible gene variants that can give rise to clinical phenotypes (20). Whole-genome sequencing has proven instrumental in revealing population- and epidemiological-level insights into pathogen detection and emergence. For example, the 2011 outbreak of the Shiga-toxin producing enteroaggregative *E. coli* O104:H4 resulted in over 3,000 infections and more than 50 deaths – rapid, open-access whole-genome sequencing analysis revealed the phylogenetic relationships between this strain and 40 previously published pathogen genomes (21). These analyses conclusively demonstrated that O104:H4's virulence was attributable to the horizontal acquisition of *stx2*, along with other unexpected traits heretofore unseen in this lineage (22). Indeed, whole-genome sequencing enables insights into a pathogen's plasticity and facilitates real-time epidemiological tracing (23).

Whole-genome sequencing has spurred the development of advanced computational techniques capable of inferring meaningful biological relationships. Advances in mathematical modeling and machine learning are now, for the first time, enabling the direct identification of antibiotic resistance determinants from the genomes of clinical isolates in as *Staphylococcus aureus*, *P. aeruginosa*, and *E. coli* (24). Moreover, mathematical modeling and high-throughput sequencing approaches have revealed that sub-inhibitory selection and step-wise adaptation play just as important a role in antibiotic treatment failure as canonical antibiotic resistance mechanisms (25). Indeed, clinical isolates from patients with relapsed *Mycobacterium tuberculosis* infection exhibit sub-breakpoint minimum inhibitory concentrations (MICs) in comparison to strains from patients durably cured (26). Mutations responsible for such subtle cellular phenotypes are readily overlooked using previous methods. Additionally, machine learning can complement traditional culture-based methods and enable the direct prediction of pathogen MICs (27, 28) and provide experimentally testable insights into antibiotic mechanisms of action (9).

## ANTIBIOTIC TREATMENT FAILURE MECHANISMS IN CLINICAL PATHOGENS

Antibiotic treatment failure is conventionally understood to be fully explained by antibiotic resistance, in which a pathogen acquires a genetic mutation either to reduce the ability of an antibiotic to inhibit its target or reduce the effective intracellular concentration of an antibiotic (15, 29). Indeed, antibiotic resistance mutations from sequenced clinical isolates frequently appear in either the target of the antibiotic, modifying the ability of an antibiotic to bind, or in the promoter regions of drug efflux pumps, inducing antibiotic export (30). Other antibiotic resistance alleles, such as genes encoding β-lactamases, commonly appear in mobile genetic elements and can become exchanged by horizontal gene transfer (31).

However, in recent years there has been a growing recognition that alternative bacterial phenotypes, such as antibiotic tolerance (in which isogenic bacteria exhibit slower killing by an antibiotic) and antibiotic persistence (in which isogenic bacteria exhibit a shallower antibiotic killing plateau), also lead to treatment failure and relapsed infection (32). Additionally, there is growing appreciation that the local microenvironment of infection can act on several aspects of bacterial physiology to alter antibiotic treatment efficacy (33, 34). In fact, the local metabolic microenvironment of an infection is highly dynamic and local metabolites induced by either infection or antibiotic treatment itself can inhibit a pathogen's cellular response to antibiotic exposure (35).

It is clear that antibiotic-target interactions alone are insufficient for explaining antibiotic treatment failure in human patients. To address these knowledge gaps, interpretable machine learning approaches are being developed, which seek to

rapidly generate experimentally testable hypotheses for biological phenomena. In one of the earliest demonstrations of these, a biochemical screen was performed to measure changes in antibiotic efficacy following metabolic stimulation, and genome-scale metabolic modeling simulations were performed to estimate metabolic reaction activities in each screening condition (**Figure 1A**). By applying machine learning to these data, purine biosynthesis was identified as a prominent player that governs antibiotic efficacy (9), highlighting a target-independent aspect of bacterial physiology is commonly involved in the lethal process of diverse bactericidal antibiotics. In light of the central role that purine metabolites also play in regulating the immune system (36), these results are also suggestive of mechanisms by which the patient-specific metabolic environment of an infection can promote drug tolerance or antibiotic treatment failure.

In another study, a metabolic model-based machine learning classifier was developed, which uses flux balance analysis to estimate the biochemical effects of genetic mutations characterized from clinical isolates (**Figure 1B**). Applying this approach to a large collection of genomes from drug-tested *M. tuberculosis* strains, novel metabolic resistance mechanisms to first-line tuberculosis antibiotics were discovered (10). These two examples illustrate how network models can serve as quantitative knowledgebases (37) and be combined with machine learning analyses to learn molecular mechanisms responsible for antibiotic treatment failures directly from clinical isolates (38).

## NUCLEOTIDE METABOLISM IN ANTIBIOTIC TREATMENT FAILURE

Bacterial metabolism is now understood to be an important physiological regulator of antibiotic efficacy (39). Across living systems, cellular metabolism is governed by the synthesis, allocation, and utilization of energy; and a growing number of studies demonstrate that metabolic dormancy protects cells from antibiotic treatment by inducing a phenotypically tolerant physiological state (29). Moreover, ATP synthesis correlates with the lethality of bactericidal antibiotics better than bacterial growth rates (40), suggesting that antibiotic-induced lethality is an active process and not merely a passive consequence of the loss-of-function of an essential gene product.

In particular, bactericidal antibiotics have been shown to elevate central carbon metabolism activity (41, 42) and trigger the formation of byproduct reactive oxygen species (43, 44), which damage DNA and cause bacterial lethality (45–47). These phenomena are not restricted to antibiotics, as reactive oxygen species also actively contribute to the lethality of bacterial secretase dysfunction (48) and thymine depletion (49). Moreover, defects in central carbon metabolism activity are linked to antibiotic tolerance and persistence across many bacterial species (50–53) and can be stimulated to enhance antibiotic efficacy (54, 55). However, antibiotic treatment perturbs several aspects of bacterial metabolism beyond central carbon metabolism (56), highlighting important knowledge gaps in understanding how different metabolic pathways may contribute to antibiotic treatment failure.

It may come to no surprise that nucleotide metabolism is actively involved in antibiotic efficacy (9). Nucleotides are essential metabolites and are ubiquitous to all living cells; in addition to their roles as fundamental building blocks for DNA and RNA molecules, constituting more than 20% of cellular biomass (57), nucleobases also form the molecular basis of primary energy currencies such as ATP and NADH, and many coenzymes are derived from nucleobase monomers. In fact, the thermodynamic properties of nucleobases are so special, that these metabolites synchronize cell biochemistry and regulate biochemical group transfers across diverse physiological processes (58). Moreover, the concentration of intracellular ATP

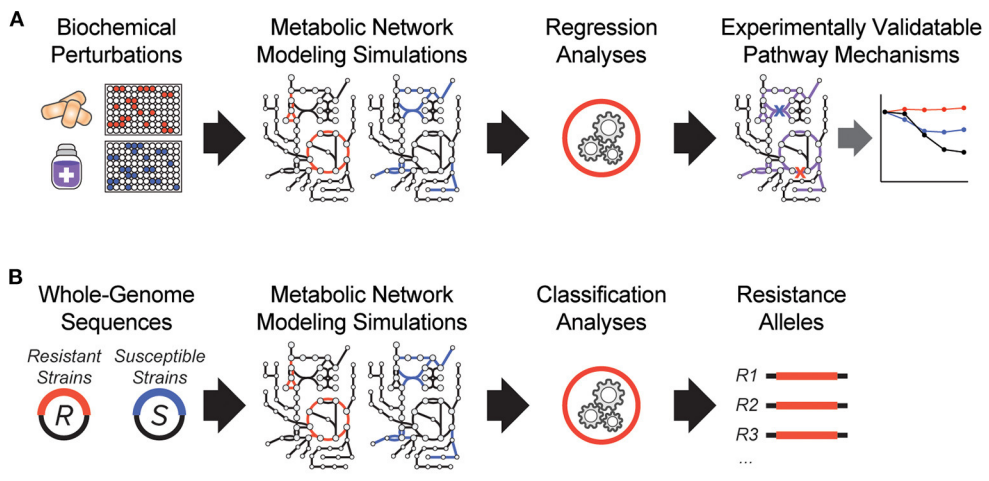



**FIGURE 1 |** Recent innovations in interpretable machine learning for studying antibiotic treatment failure. **(A)** A biochemical screen was combined with metabolic network modeling and machine learning regression analyses to elucidate pathway mechanisms of antibiotic lethality. This led to the discovery that purine biosynthesis is a critical component of bactericidal antibiotic lethality (9). **(B)** Whole-genome sequencing data from antibiotic resistant (R) and susceptible (S) strains from clinical strains were applied as modeling constraints to genome-scale metabolic models. Machine learning classification analyses were applied (10).

is tightly regulated across the tree of life and heavily buffered across environmental conditions (59).

*De novo* nucleotide biosynthesis from carbohydrates begins with the pentose phosphate pathway, which supplies phosphoribosyl pyrophosphate (prpp) as a shared substrate to the purine and pyrimidine biosynthesis pathways (**Figure 2**). These pathways produce nucleotide triphosphates which can be incorporated into DNA and RNA or processed into energy currencies that can power virtually all other biochemical processes in the cell. Interestingly, nucleotide biosynthesis is itself an energetically demanding process, costing a cell 8 ATP molecules to synthesize one adenine molecule from one glucose molecule. Indeed, cells employ a multitude of strategies to manage these tradeoffs, including prioritized nutrient usage, maintenance metabolism, and nucleotide salvage.

Antibiotic treatment imposes additional layers of complexity on these processes; cells must expend energy to mount defensive stress responses, and many antibiotics preferentially kill metabolically active cells. Specific components of nucleotide metabolism have been shown to contribute to antibiotic efficacy and protection both *in vitro* and *in vivo*. In many cases, defects in nucleotide biosynthesis have been shown to induce antibiotic persistence, suggesting these may represent a key metabolic strategy for evading antibiotic efficacy. For example, several chemogenomic screens identify nucleotide biosynthesis genes, as well as global regulators of nucleotide metabolism, as important regulators of antibiotic tolerance (60, 61). Likewise, antibiotic drug screening under nutrient limitation identified several compounds that interfere in core or peripheral nucleotide metabolism branching points (62).

Of note, purine biosynthesis frequently emerges as a key pathway responsible for antibiotic efficacy. For example, in an antibiotic persistence screen using a *S. aureus* transposon mutant library, 29% of all depleted genes were related to cellular metabolism, and of these, five were involved in purine biosynthesis (63). These *ex vivo* observations are important for understanding clinical antibiotic treatment failure, as methicillin-resistant *S. aureus* clones isolated from patients enduring multi-drug antibiotic treatment were found to possess mutations in *purR*, a transcriptional repressor of purine synthesis, within 1

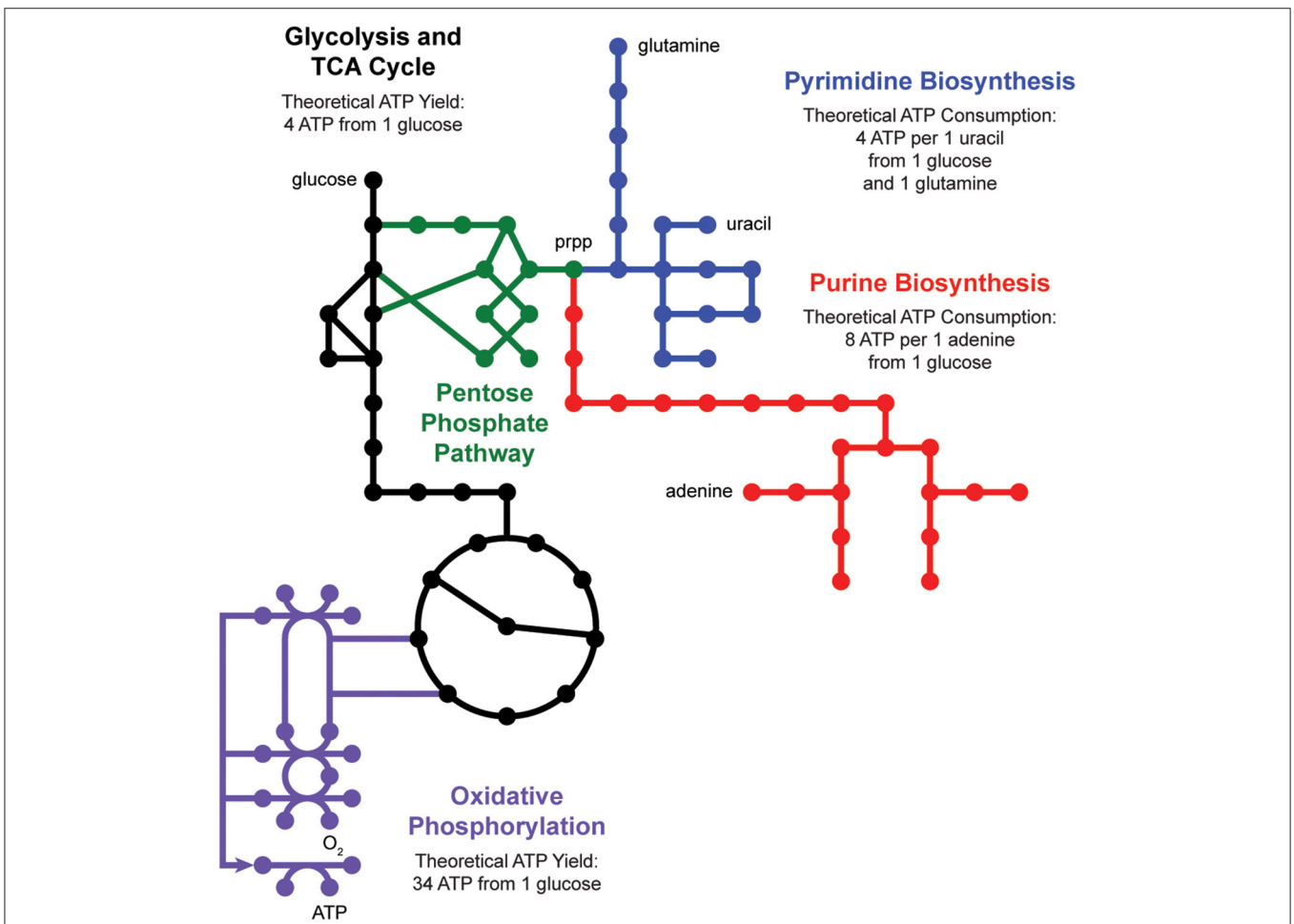



**FIGURE 2 |** Nucleotide metabolism is energetically expensive. *De novo* biosynthesis of purines (red) and pyrimidines (blue) begins with the pentose phosphate pathway (green), which generates phosphoribosyl pyrophosphate (prpp) from glycolysis (black). The energetic demand for ATP molecules to power purine and pyrimidine biosynthesis drives activity through the tricarboxylic acid (TCA) cycle (black) and oxidative phosphorylation (purple).

week of treatment. *In vitro* follow-up experiments confirmed that this mutation reduced the rate of vancomycin-induced killing, revealing the evolution of antibiotic tolerance *in vivo* (64). Importantly, this mutation preceded the onset of canonical resistance evolution; these and other studies suggest that mutations in nucleotide metabolism may help create a reservoir of pathogen cells primed to subsequently evolve target-specific antibiotic resistance alleles.

Recent microbiological studies are beginning to clarify how nucleotide metabolism contributes to antibiotic efficacy (**Figure 3**). Interpretable machine learning analyses reveals that several metabolic pathways proximal to purine biosynthesis contribute to the lethality of bactericidal antibiotics in *E. coli* (9). Purine biosynthesis becomes induced by bactericidal stress-induced adenine limitation, which can be directly measured by targeted metabolomics (56). Consequently, oxidative phosphorylation becomes elevated to meet the increased energetic demand of enhanced purine biosynthesis, increasing cellular respiration and central carbon metabolism and providing substrates for toxic reactive oxygen species (42, 43). Indeed, regulation of nucleotide metabolism appears to be a well-conserved mechanism that bacteria have evolved to handle diverse stresses (65).

Consistent with these, purine nucleotides such as (p)ppGpp function as universal alarmones for transcriptionally activating the stringent response and other bacterial stress responses as evolutionally conserved strategies for surviving nutrient limitation and other environmental stressors (66, 67). Intracellular accumulation of (p)ppGpp and related purine alarmones can induce antibiotic tolerance by promoting growth arrest (68) and entry to antibiotic persister states (69). Recent studies demonstrate that in additional to these transcriptionally mediated programs, (p)ppGpp can also inhibit nucleotide metabolism directly by binding several enzymes involved in purine biosynthesis, including PurF and Gsk (70, 71). These data collectively support a central role for nucleotide metabolism in antibiotic treatment efficacy.

It is interesting to note that nucleotide metabolism is also very important for the *in vivo* pathogenesis of diverse bacterial infections, and may be required for a pathogen's growth and survival within the host environment (72). For instance, *S. aureus* cells with transposon insertions in *purB* fail to establish bone infections in mice (73) and deletion of purine biosynthesis genes prevents uropathogenic *E. coli* from expanding into intracellular bladder epithelial cells (74). Likewise, *in vivo* studies of methicillin-resistant *S. aureus* showed that purine

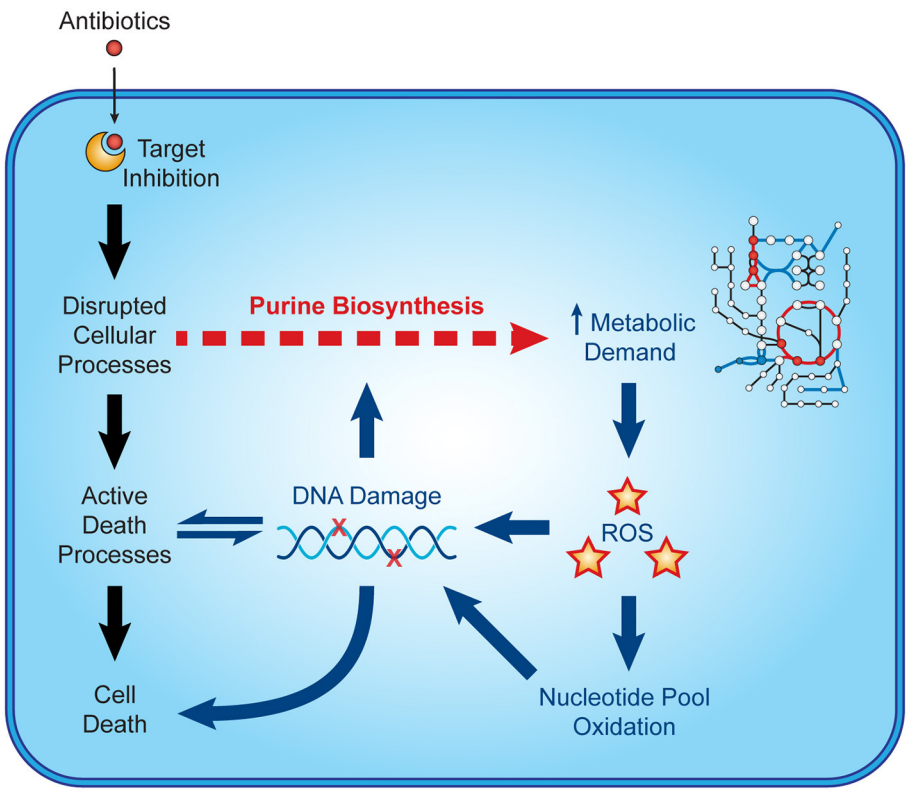



**FIGURE 3 |** Nucleotide metabolism contributes to antibiotic lethality. In addition to their target-specific effects, bactericidal antibiotics induce purine biosynthesis, which increases activity in central metabolism. Increases in central metabolism stimulate the production of toxic reactive oxygen species, which oxidize nucleotides and damage DNA. These insults to DNA and the nucleotide pool induce bacterial death and may further potentiate purine biosynthesis.

biosynthesis was causally linked to survival during endovascular infection (11). Collectively, it is clear that nucleotide metabolism, particularly purine biosynthesis, plays an important role in bacterial pathogenesis and in the response to antibiotic stress.

## DISCUSSION

The growing challenge of clinical antibiotic failure demands renewed attention into the study of antibiotic mechanisms of action and the discovery of new antimicrobial compounds. Digital technologies such as whole-genome sequencing, machine learning, mass spectrometry and predictive modeling are likely to transform the clinical management of bacterial infections in the coming decades. Exciting developments in machine learning are, for the first time, enabling the rapid discovery of novel classes of antimicrobial compounds (75) and the rapid identification of bacterial pathogens in the clinic (5). Advances in mass spectrometry-based metabolomics are enabling the rapid discovery of antimicrobial mechanisms of action (76). Advances in predictive modeling (7) are enabling new understanding into the complex ecology of microbial communities (77).

The discovery that nucleotide metabolism is involved in antibiotic efficacy has several translational implications. Unlike the Mueller-Hinton or Luria-Bertani media commonly used by clinical and academic microbiology laboratories, the metabolic microenvironment of a bacterial infection is dynamically enriched for nucleotide metabolites during infection (35). In fact, purine metabolites are important regulators of innate immunity (36), playing dual roles in regulating the host response to infection and the pathogen response to antibiotics. Nucleotide analogs are also commonly used to treat human cancers and viral infections and have potential to address antimicrobial resistance in the clinic (78, 79).

Nucleotide metabolism is one of the oldest areas of bacterial physiology to be investigated, with early studies into bacterial purine and pyrimidine metabolism predating the discovery of the lac operon (80, 81). Interest in nucleotide metabolism is mounting a resurgence, spurred by the growing recognition that nucleotides play important roles in both immunometabolism (82, 83) and cancer pathogenesis (84). Given that purine and pyrimidines exert opposing effects on antibiotic efficacy and carbon metabolism in bacteria (9), nucleotide metabolism represents an exciting open frontier for future studies in bacterial physiology and antibiotic treatment failure.

Concurrently, new digitalization techniques are becoming increasingly democratized and are poised to transform our basic and translational understanding of how nucleotide metabolism may contribute to antibiotic efficacy. Advances in predictive modeling (7) and non-targeted metabolomics (85) are revealing the diverse systems-level consequences of antibiotic stress. Quantitative microscopy advances (86) are enabling detection of antibiotic tolerance and resistance at single-cell resolution. Advances in transposon insertion sequencing (87) and adaptive lab evolution (88) are revealing new mechanisms for antibiotic resistance. Indeed, it would be exciting for future discoveries to reveal how nucleotide metabolism may contribute to antibiotic failure mechanisms beyond persistence (11) and potentially rewrite our understanding of antimicrobial resistance (29).

## AUTHOR CONTRIBUTIONS

AL and JY planned, wrote, and edited the manuscript. All authors critically read, reviewed, and approved the final version of the manuscript.

## FUNDING

This work was supported by grant R00-GM118907 from the National Institutes of Health to JY.

## REFERENCES

1. World Health Organization. *Antimicrobial Resistance: Global Report on Surveillance*. Geneva (2014).
2. Brown ED, Wright GD. Antibacterial drug discovery in the resistance era. *Nature.* (2016) 529:336–43. doi: 10.1038/nature17042
3. Buchan BW, Ledeboer NA. Emerging technologies for the clinical microbiology laboratory. *Clin Microbiol Rev.* (2014) 27:783–822. doi: 10.1128/CMR.00003-14
4. Han D, Li Z, Li R, Tan P, Zhang R, Li J. mNGS in clinical microbiology laboratories: on the road to maturity. *Crit Rev Microbiol.* (2019) 45:668–85. doi: 10.1080/1040841X.2019.1681933
5. Smith KP, Kirby JE. Image analysis and artificial intelligence in infectious disease diagnostics. *Clin Microbiol Infect.* (2020) 26:1318–23. doi: 10.1016/j.cmi.2020.03.012
6. O'brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. *Cell.* (2015) 161:971–87. doi: 10.1016/j.cell.2015.05.019
7. Lopatkin AJ, Collins JJ. Predictive biology: modelling, understanding and harnessing microbial complexity. *Nat Rev Microbiol.* (2020) 18:507–20. doi: 10.1038/s41579-020-0372-5
8. Yu MK, Ma J, Fisher J, Kreisberg JF, Raphael BJ, Ideker T. Visible machine learning for biomedicine. *Cell.* (2018) 173:1562–5. doi: 10.1016/j.cell.2018.05.056
9. Yang JH, Wright SN, Hamblin M, Mccloskey D, Alcantar MA, Schrubbers L, et al. A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell.* (2019) 177:1649–61.e1649. doi: 10.1016/j.cell.2019.04.016
10. Kavvas ES, Yang L, Monk JM, Heckmann D, Palsson BO. A biochemically-interpretable machine learning classifier for microbial GWAS. *Nat Commun.* (2020) 11:2580. doi: 10.1038/s41467-020-16310-9
11. Li L, Abdelhady W, Donegan NP, Seidl K, Cheung A, Zhou YF, et al. Role of purine biosynthesis in persistent methicillin-resistant *Staphylococcus aureus* (MRSA) infection. *J Infect Dis.* (2018) 218:1367–77. doi: 10.1093/infdis/jiy340
12. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet.* (2012) 13:601–12. doi: 10.1038/nrg3226
13. Levy SB, Mcmurry L. Plasmid-determined tetracycline resistance involves new transport systems for tetracycline. *Nature.* (1978) 276:90–2. doi: 10.1038/276090a0
14. Ball PR, Shales SW, Chopra I. Plasmid-mediated tetracycline resistance in *Escherichia coli* involves increased efflux of the antibiotic. *Biochem Biophys Res Commun.* (1980) 93:74–81. doi: 10.1016/S0006-291X(80)80247-6
15. Blair JM, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJ. Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol.* (2015) 13:42–51. doi: 10.1038/nrmicro3380

16. Hughes D, Andersson DI. Evolutionary trajectories to antibiotic resistance. *Annu Rev Microbiol.* (2017) 71:579–96. doi: 10.1146/annurev-micro-090816-093813
17. Lopatkin AJ, Bening SC, Manson AL, Stokes JM, Kohanski MA, Badran AH, et al. Clinically relevant mutations in core metabolic genes confer antibiotic resistance. *Science.* (2021) 371:eaba0862. doi: 10.1126/science.aba0862
18. Lazar V, Nagy I, Spohn R, Csorgo B, Gyorkei A, Nyerges A, et al. Genome-wide analysis captures the determinants of the antibiotic cross-resistance interaction network. *Nat Commun.* (2014) 5:4352. doi: 10.1038/ncomms5352
19. Baym M, Lieberman TD, Kelsic ED, Chait R, Gross R, Yelin I, et al. Spatiotemporal microbial evolution on antibiotic landscapes. *Science.* (2016) 353:1147–51. doi: 10.1126/science.aag0822
20. Anahtar MN, Yang JH, Kanjilal S. Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *J Clin Microbiol.* (2021). doi: 10.1128/JCM.01260-20 [Epub ahead of print].
21. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med.* (2011) 365:709–17. doi: 10.1056/NEJMoa1106920
22. Karch H, Denamur E, Dobrindt U, Finlay BB, Hengge R, Johannes L, et al. The enemy within us: lessons from the 2011 European *Escherichia coli* O104:H4 outbreak. *EMBO Mol Med.* (2012) 4:841–8. doi: 10.1002/emmm.201201662
23. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics.* (2018) 34:4121–3. doi: 10.1093/bioinformatics/bty407
24. Hyun JC, Kavvas ES, Monk JM, Palsson BO. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. *PLoS Comput Biol.* (2020) 16:e1007608. doi: 10.1371/journal.pcbi.1007608
25. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, Garcia-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. *J Biotechnol.* (2017) 243:16–24. doi: 10.1016/j.jbiotec.2016.12.022
26. Colangeli R, Jedrey H, Kim S, Connell R, Ma S, Chippada Venkata UD, et al. Bacterial factors that predict relapse after tuberculosis therapy. *N Engl J Med.* (2018) 379:823–33. doi: 10.1056/NEJMoa1715849
27. Monk JM. Predicting antimicrobial resistance and associated genomic features from whole-genome sequencing. *J Clin Microbiol.* (2019) 57:e01610–8. doi: 10.1128/JCM.01610-18
28. Nguyen M, Long SW, Mcdermott PF, Olsen RJ, Olson R, Stevens RL, et al. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal salmonella. *J Clin Microbiol.* (2019) 57:e01260–18. doi: 10.1128/JCM.01260-18
29. Schrader SM, Vaubourgeix J, Nathan C. Biology of antimicrobial resistance and approaches to combat it. *Sci Translat Med.* (2020) 12:eaaz6992. doi: 10.1126/scitranslmed.aaz6992
30. Mcarthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother.* (2013) 57:3348–57. doi: 10.1128/AAC.00419-13
31. Maiden MC. Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacteria. *Clin Infect Dis.* (1998) 27(Suppl. 1):S12–20. doi: 10.1086/514917
32. Brauner A, Fridman O, Gefen O, Balaban NQ. Distinguishing between resistance, tolerance and persistence to antibiotic treatment. *Nat Rev Microbiol.* (2016) 14:320–30. doi: 10.1038/nrmicro.2016.34
33. Yang JH, Bening SC, Collins JJ. Antibiotic efficacy-context matters. *Curr Opin Microbiol.* (2017) 39:73–80. doi: 10.1016/j.mib.2017.09.002
34. Radlinski L, Conlon BP. Antibiotic efficacy in the complex infection environment. *Curr Opin Microbiol.* (2018) 42:19–24. doi: 10.1016/j.mib.2017.09.007
35. Yang JH, Bhargava P, Mccloskey D, Mao N, Palsson BO, Collins JJ. Antibiotic-induced changes to the host metabolic environment inhibit drug efficacy and alter immune function. *Cell Host Microbe.* (2017) 22:757–65.e753. doi: 10.1016/j.chom.2017.10.020
36. Cekic C, Linden J. Purinergic regulation of the immune system. *Nat Rev Immunol.* (2016) 16:177–92. doi: 10.1038/nri.2016.4
37. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, et al. iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat Biotechnol.* (2017) 35:904–8. doi: 10.1038/nbt.3956
38. Hicks ND, Yang J, Zhang X, Zhao B, Grad YH, Liu L, et al. Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. *Nat Microbiol.* (2018) 3:1032–42. doi: 10.1038/s41564-018-0218-3
39. Stokes JM, Lopatkin AJ, Lobritz MA, Collins JJ. Bacterial metabolism and antibiotic efficacy. *Cell Metab.* (2019) 30:251–9. doi: 10.1016/j.cmet.2019.06.009
40. Lopatkin AJ, Stokes JM, Zheng EJ, Yang JH, Takahashi MK, You L, et al. Bacterial metabolic state more accurately predicts antibiotic lethality than growth rate. *Nat Microbiol.* (2019) 4:2109–17. doi: 10.1038/s41564-019-0536-0
41. Dwyer DJ, Belenky PA, Yang JH, Macdonald IC, Martell JD, Takahashi N, et al. Antibiotics induce redox-related physiological alterations as part of their lethality. *Proc Natl Acad Sci USA.* (2014) 111:E2100–9. doi: 10.1073/pnas.1401876111
42. Lobritz MA, Belenky P, Porter CB, Gutierrez A, Yang JH, Schwarz EG, et al. Antibiotic efficacy is linked to bacterial cellular respiration. *Proc Natl Acad Sci USA.* (2015) 112:8173–80. doi: 10.1073/pnas.1509743112
43. Kohanski MA, Dwyer DJ, Hayete B, Lawrence CA, Collins JJ. A common mechanism of cellular death induced by bactericidal antibiotics. *Cell.* (2007) 130:797–810. doi: 10.1016/j.cell.2007.06.049
44. Zhao X, Drlica K. Reactive oxygen species and the bacterial response to lethal stress. *Curr Opin Microbiol.* (2014) 21:1–6. doi: 10.1016/j.mib.2014.06.008
45. Foti JJ, Devadoss B, Winkler JA, Collins JJ, Walker GC. Oxidation of the guanine nucleotide pool underlies cell death by bactericidal antibiotics. *Science.* (2012) 336:315–9. doi: 10.1126/science.1219192
46. Fan X.-Y, Tang B.-K, Xu Y.-Y, Han A.-X, Shi K.-X, Wu Y.-K, et al. Oxidation of dCTP contributes to antibiotic lethality in stationary-phase mycobacteria. *Proc Natl Acad Sci USA.* (2018) 115:2210–5. doi: 10.1073/pnas.1719627115
47. Hong Y, Zeng J, Wang X, Drlica K, Zhao X. Post-stress bacterial cell death mediated by reactive oxygen species. *Proc Natl Acad Sci USA.* (2019) 116:10064–71. doi: 10.1073/pnas.1901730116
48. Takahashi N, Gruber CC, Yang JH, Liu X, Braff D, Yashaswini C, et al. Lethality of MalE-LacZ hybrid protein shares mechanistic attributes with oxidative component of antibiotic lethality. *Proc Natl Acad Sci USA.* (2017) 114:9164–9. doi: 10.1073/pnas.1707466114
49. Hong Y, Li L, Luan G, Drlica K, Zhao X. Contribution of reactive oxygen species to thymineless death in *Escherichia coli*. *Nat Microbiol.* (2017) 2:1667–75. doi: 10.1038/s41564-017-0037-y
50. Baek SH, Li AH, Sassetti CM. Metabolic regulation of mycobacterial growth and antibiotic sensitivity. *PLoS Biol.* (2011) 9:e1001065. doi: 10.1371/journal.pbio.1001065
51. Shan Y, Brown Gandt A, Rowe SE, Deisinger JP, Conlon BP, Lewis K. ATP-dependent persister formation in *Escherichia coli*. *MBio.* (2017) 8:e02267–16. doi: 10.1128/mBio.02267-16
52. Lee JJ, Lee SK, Song N, Nathan TO, Swarts BM, Eum SY, et al. Transient drug-tolerance and permanent drug-resistance rely on the trehalose-catalytic shift in *Mycobacterium tuberculosis*. *Nat Commun.* (2019) 10:2928. doi: 10.1038/s41467-019-10975-7
53. Zalis EA, Nuxoll AS, Manuse S, Clair G, Radlinski LC, Conlon BP, et al. Stochastic variation in expression of the tricarboxylic acid cycle produces persister cells. *MBio.* (2019) 10:e01930–19. doi: 10.1128/mBio.01930-19
54. Gutierrez A, Jain S., Bhargava P, Hamblin M, Lobritz MA, Collins JJ. Understanding and sensitizing density-dependent persistence to quinolone antibiotics. *Mol Cell.* (2017) 68:1147–54.e1143. doi: 10.1016/j.molcel.2017.11.012
55. Meylan S, Porter CBM, Yang JH, Belenky P, Gutierrez A, Lobritz MA, et al. Carbon sources tune antibiotic susceptibility in *Pseudomonas aeruginosa* via tricarboxylic acid cycle control. *Cell Chem Biol.* (2017) 24:195–206. doi: 10.1016/j.chembiol.2016.12.015
56. Belenky P, Ye JD, Porter CB, Cohen NR, Lobritz MA, Ferrante T, et al. Bactericidal antibiotics induce toxic metabolic perturbations that lead to cellular damage. *Cell Rep.* (2015) 13:968–80. doi: 10.1016/j.celrep.2015.09.059
57. Milo R, Phillips R. *Cell Biology by the Numbers*. New York, NY: Garland Science, Taylor & Francis Group (2016).

58. Walsh CT, Tu BP, Tang Y. Eight kinetically stable but thermodynamically activated molecules that power cell metabolism. *Chem Rev.* (2018) 118:1460–94. doi: 10.1021/acs.chemrev.7b00510
59. Chapman AG, Atkinson DE. Adenine nucleotide concentrations and turnover rates. Their correlation with biological activity in bacteria and yeast. *Adv Microb Physiol.* (1977) 15:253–306. doi: 10.1016/S0065-2911(08)60318-5
60. Hansen S, Lewis K, Vulic M. Role of global regulators and nucleotide metabolism in antibiotic tolerance in *Escherichia coli*. *Antimicrob Agents Chemother.* (2008) 52:2718–26. doi: 10.1128/AAC.00144-08
61. Stokes JM, Gutierrez A, Lopatkin AJ, Andrews IW, French S, Matic I, et al. A multiplexable assay for screening antibiotic lethality against drug-tolerant bacteria. *Nat Methods.* (2019) 16:303–6. doi: 10.1038/s41592-019-0333-y
62. Zlitni S, Ferruccio LF, Brown ED. Metabolic suppression identifies new antibacterial inhibitors under nutrient limitation. *Nat Chem Biol.* (2013) 9:796–804. doi: 10.1038/nchembio.1361
63. Yee R, Cui P, Shi W, Feng J, Zhang Y. Genetic screen reveals the role of purine metabolism in *Staphylococcus aureus* persistence to rifampicin. *Antibiotics.* (2015) 4:627–42. doi: 10.3390/antibiotics4040627
64. Liu J, Gefen O, Ronin I, Bar-Meir M, Balaban NQ. Effect of tolerance on the evolution of antibiotic resistance under drug combinations. *Science.* (2020) 367:200–4. doi: 10.1126/science.aay3041
65. Fitzsimmons LF, Liu L, Kim JS, Jones-Carson J, Vazquez-Torres A. Salmonella reprograms nucleotide metabolism in its adaptation to nitrosative stress. *MBio.* (2018) 9:e002118. doi: 10.1128/mBio.00211-18
66. Hauryliuk V, Atkinson GC, Murakami KS, Tenson T, Gerdes K. Recent functional insights into the role of (p)ppGpp in bacterial physiology. *Nat Rev Microbiol.* (2015) 13:298–309. doi: 10.1038/nrmicro3448
67. Irving SE, Choudhury NR, Corrigan RM. The stringent response and physiological roles of (pp)pGpp in bacteria. *Nat Rev Microbiol.* (2020). doi: 10.1038/s41579-020-00470-y
68. Spira B, Ospino K. Diversity in *E. coli* (p)ppGpp levels and its consequences. *Front Microbiol.* (2020) 11:1759. doi: 10.3389/fmicb.2020.01759
69. Amato SM, Orman MA, Brynildsen MP. Metabolic control of persister formation in *Escherichia coli*. *Mol Cell.* (2013) 50:475–87. doi: 10.1016/j.molcel.2013.04.002
70. Wang B, Dai P, Ding D, Del Rosario A, Grant RA, Pentelute BL, et al. Affinity-based capture and identification of protein effectors of the growth regulator ppGpp. *Nat Chem Biol.* (2019) 15:141–50. doi: 10.1038/s41589-018-0183-4
71. Wang B, Grant RA, Laub MT. ppGpp coordinates nucleotide and amino-acid synthesis in *E. coli* during starvation molecular. *Cell.* (2020) 80:29–42.e10. doi: 10.1016/j.molcel.2020.08.005
72. Samant S, Lee H, Ghassemi M, Chen J, Cook JL, Mankin AS, et al. Nucleotide biosynthesis is critical for growth of bacteria in human blood. *PLoS Pathog.* (2008) 4:e37. doi: 10.1371/journal.ppat.0040037
73. Potter AD, Butrico CE, Ford CA, Curry JM, Trenary IA, Tummarakota SS, et al. Host nutrient milieu drives an essential role for aspartate biosynthesis during invasive *Staphylococcus aureus* infection. *Proc Natl Acad Sci USA.* (2020) 117:12394–401. doi: 10.1073/pnas.1922211117
74. Shaffer CL, Zhang EW, Dudley AG, Dixon BREA, Guckes KR, Breland EJ, et al. purine biosynthesis metabolically constrains intracellular survival of uropathogenic *Escherichia coli*. *Infect Immun.* (2017) 85:e00471–16. doi: 10.1128/IAI.00471-16
75. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. *Cell.* (2020) 181:475–83. doi: 10.1016/j.cell.2020.04.001
76. Zampieri M, Szappanos B, Buchieri MV, Trauner A, Piazza I, Picotti P, et al. High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. *Sci Transl Med.* (2018) 10:eaal3973. doi: 10.1126/scitranslmed.aal3973
77. Meredith HR, Andreani V, Ma HR, Lopatkin AJ, Lee AJ, Anderson DJ, et al. Applying ecological resistance and resilience to dissect bacterial antibiotic responses. *Sci Adv.* (2018) 4:eaau1873. doi: 10.1126/sciadv.aau1873
78. Liu Y, Yang K, Jia Y, Shi J, Tong Z, Wang Z. Thymine sensitizes gram-negative pathogens to antibiotic killing. *Front Microbiol.* (2021) 12:622798. doi: 10.3389/fmicb.2021.622798
79. Serpi M, Ferrari V, Pertusati F. Nucleoside derived antibiotics to fight microbial drug resistance: new utilities for an established class of drugs? *J Med Chem.* (2016) 59:10343–82. doi: 10.1021/acs.jmedchem.6b00325
80. Koch AL, Putnam FW, Evans EA. The purine metabolism of *Escherichia coli*. *J Biol Chem.* (1952) 197:105–12. doi: 10.1016/S0021-9258(18)55658-1
81. Friedman S, Gots JS. The purine and pyrimidine metabolism of normal and phage-infected *Escherichia coli*. *J Biol Chem.* (1953) 201:125–35. doi: 10.1016/S0021-9258(18)71354-9
82. Mazumdar C, Driggers EM, Turka LA. The untapped opportunity and challenge of immunometabolism: a new paradigm for drug discovery. *Cell Metab.* (2020) 31:26–34. doi: 10.1016/j.cmet.2019.11.014
83. Palsson-Mcdermott EM, O'neill LJ. Targeting immunometabolism as an anti-inflammatory strategy. *Cell Res.* (2020) 30:300–14. doi: 10.1038/s41422-020-0291-z
84. Campos-Contreras ADR, Diaz-Munoz M, Vazquez-Cuevas FG. Purinergic signaling in the hallmarks of cancer. *Cells.* (2020) 9:1612. doi: 10.3390/cells9071612
85. Zampieri M, Zimmermann M, Claassen M, Sauer U. Nontargeted metabolomics reveals the multilevel response to antibiotic perturbations. *Cell Rep.* (2017) 19:1214–28. doi: 10.1016/j.celrep.2017.04.002
86. Herricks T, Donczew M, Mast FD, Rustad T, Morrison R, Sterling TR, et al. ODELAM, rapid sequence-independent detection of drug resistance in isolates of Mycobacterium tuberculosis. *Elife.* (2020) 9:e56613. doi: 10.1101/2020.03.17.995480
87. Cain AK, Barquist L, Goodman AL, Paulsen IT, Parkhill J, Van Opijnen T. A decade of advances in transposon-insertion sequencing. *Nat Rev Gene.* (2020) 21:526–40. doi: 10.1038/s41576-020-0244-x
88. Lukačišinová M, Fernando B, Bollenbach T. Highly parallel lab evolution reveals that epistasis can curb the evolution of antibiotic resistance. *Nat Commun.* (2020) 11:3105. doi: 10.1038/s41467-020-16932-z


**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2021 Lopatkin and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Medicine

SYSTEMATIC REVIEW
published: 28 May 2021
doi: 10.3389/fmed.2021.607952

# Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review

*Michael Moor*[1,2*†], *Bastian Rieck*[1,2†], *Max Horn*[1,2], *Catherine R. Jutzeler*[1,2‡] *and Karsten Borgwardt*[1,2‡]

[1] *Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, Eidgenössische Technische Hochschule Zürich (ETH Zurich), Basel, Switzerland,* [2] *SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland*

**Background:** Sepsis is among the leading causes of death in intensive care units (ICUs) worldwide and its recognition, particularly in the early stages of the disease, remains a medical challenge. The advent of an affluence of available digital health data has created a setting in which machine learning can be used for digital biomarker discovery, with the ultimate goal to advance the early recognition of sepsis.

**Objective:** To systematically review and evaluate studies employing machine learning for the prediction of sepsis in the ICU.

**Data Sources:** Using Embase, Google Scholar, PubMed/Medline, Scopus, and Web of Science, we systematically searched the existing literature for machine learning-driven sepsis onset prediction for patients in the ICU.

**Study Eligibility Criteria:** All peer-reviewed articles using machine learning for the prediction of sepsis onset in adult ICU patients were included. Studies focusing on patient populations outside the ICU were excluded.

**Study Appraisal and Synthesis Methods:** A systematic review was performed according to the PRISMA guidelines. Moreover, a quality assessment of all eligible studies was performed.

**Results:** Out of 974 identified articles, 22 and 21 met the criteria to be included in the systematic review and quality assessment, respectively. A multitude of machine learning algorithms were applied to refine the early prediction of sepsis. The quality of the studies ranged from "poor" (satisfying $\leq$ 40% of the quality criteria) to "very good" (satisfying $\geq$ 90% of the quality criteria). The majority of the studies ($n = 19$, 86.4%) employed an offline training scenario combined with a horizon evaluation, while two studies implemented an online scenario ($n = 2$, 9.1%). The massive inter-study heterogeneity in terms of model development, sepsis definition, prediction time windows, and outcomes precluded a meta-analysis. Last, only two studies provided publicly accessible source code and data sources fostering reproducibility.

**Limitations:** Articles were only eligible for inclusion when employing machine learning algorithms for the prediction of sepsis onset in the ICU. This restriction led to the exclusion of studies focusing on the prediction of septic shock, sepsis-related mortality, and patient populations outside the ICU.

OPEN ACCESS

***Edited by:***
*Belén Rodriguez-Sanchez, Gregorio Marañón Hospital, Spain*

***Reviewed by:***
*Axel Nierhaus, University of Hamburg, Germany*
*Gilbert Greub, University of Lausanne, Switzerland*

****Correspondence:***
*Michael Moor*
*michael.moor@bsse.ethz.ch*

*†These authors have contributed equally to this work*

*‡These authors jointly directed this work*

***Specialty section:***
*This article was submitted to Infectious Diseases – Surveillance, Prevention and Treatment, a section of the journal Frontiers in Medicine*

***Received:*** *18 September 2020*
***Accepted:*** *04 March 2021*
***Published:*** *28 May 2021*

***Citation:***
*Moor M, Rieck B, Horn M, Jutzeler CR and Borgwardt K (2021) Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. Front. Med. 8:607952. doi: 10.3389/fmed.2021.607952*

**Conclusions and Key Findings:** A growing number of studies employs machine learning to optimize the early prediction of sepsis through digital biomarker discovery. This review, however, highlights several shortcomings of the current approaches, including low comparability and reproducibility. Finally, we gather recommendations how these challenges can be addressed before deploying these models in prospective analyses.

**Systematic Review Registration Number:** CRD42020200133.

**Keywords: sepsis, machine learning, onset prediction, early detection, systematic review**

## 1. INTRODUCTION

Sepsis is a life-threatening organ dysfunction triggered by dysregulated host response to infection (1) and constitutes a major global health concern (2). Despite promising medical advances over the last decades, sepsis remains among the most common causes of in-hospital deaths. It is associated with an alarmingly high mortality and morbidity, and massively burdens the health care systems world-wide (2–5). In parts, this can be attributed to challenges related to early recognition of sepsis and initiation of timely and appropriate treatment (6). A growing number of studies suggests that the mortality increases with every hour the antimicrobial intervention is delayed, further underscoring the importance of timely recognition and initiation of treatment (6–8). A major challenge to early recognition is to distinguish sepsis from disease states (e.g., inflammation) that are hallmarked by similar clinical signs (e.g., change in vitals), symptoms (e.g., fever), and molecular manifestations (e.g., dysregulated host response) (9, 10). Owing to the systemic nature of sepsis, biological and molecular correlates—also known as biomarkers—have been proposed to refine the diagnosis and detection of sepsis (5). However, despite considerable efforts to identify suitable biomarkers, there is yet no single biomarker or set thereof that is universally accepted for sepsis diagnosis and treatment, mainly due to the lack of sensitivity and specificity (11, 12).

In addition to the conventional approaches, *data-driven biomarker discovery* has gained momentum over the last decades and holds the promise to overcome existing hurdles. The goal of this approach is to mine and exploit health data with quantitative computational approaches, such as machine learning. An ever-increasing amount of data, including laboratory, vital, genetic, molecular, as well as clinical data and health history, is available in digital form and at high resolution for individuals at risk and for patients suffering from sepsis (13). This versatility of the data allows to search for digital biomarkers in a holistic fashion as opposed to a reductionist approach (e.g., solely focusing on hematological markers). Machine learning models can naturally handle the wealth and complexity of digital patient data by learning predictive patterns in the data, which in turn can be used to make accurate predictions about which patient is developing sepsis (14, 15). Searching predictive patterns is conventionally done either in a supervised or unsupervised fashion. Supervised learning refers to algorithms that learn from labeled training data (e.g., patients have sepsis or not) to predict outcomes for unforeseen data. In contrast, in unsupervised learning, the data have no labels and the algorithm detects (known and unknown) patterns based on the data provided. Over the last decades, multiple studies have successfully employed a variety of computational models to tackle the challenge of predicting sepsis at the earliest time point possible (16–18). For instance, Futoma et al. proposed to combine multi-task Gaussian processes imputation together with a recurrent neural network in one end-to-end trainable framework (multi-task Gaussian process recurrent neural network [MGP-RNN]). They were able to predict sepsis 17 h prior to the first administration of antibiotics and 36 h before a definition for sepsis was met (19). This strategy was motivated by Li and Marlin (20), who first proposed the so-called Gaussian process adapter that combines single-task Gaussian processes imputation with neural networks in an end-to-end learning setting. A more recent study further improved predictive performance by combining the Gaussian process adapter framework with temporal convolutional networks (MGP-TCN) as well as leveraging a dynamic time warping approach for the early prediction of sepsis (21).

Considering the rapid pace at which the research in this field is moving forward, it is important to summarize and critically assess the state of the art. Thus, the aim of this review was to provide a comprehensive overview of the current state of machine learning models that have been employed for the search of digital biomarkers to aid the early prediction of sepsis in the intensive care unit (ICU). To this end, we systematically reviewed the literature and performed a quality assessment of all eligible studies. Based on our findings, we also provide some recommendations for forthcoming studies that plan to use machine learning models for the early prediction of sepsis.

## 2. METHODS

The study protocol was registered with and approved by the international prospective register of systematic reviews (PROSPERO) before the start of the study (registration number: CRD42020200133). We followed the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement (22).

## 2.1. Search Strategy and Selection Criteria

Five bibliographic databases were systematically searched, i.e., EMBASE, Google Scholar, PubMed/Medline, Scopus, and Web of Science, using the time range from their respective inception dates to July 20, 2020. Google Scholar was searched using the tool "Publish or Perish" (version 7.23.2852.7498) (23). Our search was not restricted by language. The search term string was constructed as `("sepsis prediction" OR "sepsis detection") AND ("machine learning" OR "artificial intelligence")` to include publications focusing on (early) onset prediction of sepsis with different machine learning methods. The full search strategy is provided in **Supplementary Table 1**.

## 2.2. Selection of Studies

Two investigators (MM and CRJ) independently screened the titles, abstracts, and full texts retrieved from Google Scholar in order to determine the eligibility of the studies. Google Scholar was selected by virtue of its promise of an inclusive query that also captures conference proceedings, which are highly relevant to the field of machine learning but not necessarily indexed by other databases. In a second step, two investigators (MM and MH) queried EMBASE, PubMed, Scopus, and Web of Science for additional studies. Eligibility criteria were also applied to the full-text articles during the final selection. In case multiple articles reported on a single study, the article that provided the most data and details was selected for further synthesis. We quantified the inter-rater agreement for study selection using Cohen's kappa ($\kappa$) coefficient (24). All disagreements were discussed and resolved at a consensus meeting.

## 2.3. Inclusion and Exclusion Criteria

All full-text, peer-reviewed articles[1] using machine learning for the prediction of sepsis onset in the ICU were included. Although the 2016 consensus statement abandoned the term "severe sepsis" (1), studies published prior to the revised consensus statement targeting severe sepsis were also included in our review. Furthermore, to be included, studies must have provided sufficient information on the machine learning algorithms used for the analysis, definition of sepsis (e.g., Sepsis-3), and sepsis onset definition (e.g., time of suspicion of infection). We excluded duplicates, non-peer reviewed articles (e.g., preprints), reviews, meta-analyses, abstracts, editorials, commentaries, perspectives, patents, letters with insufficient data, studies on non-human species and children/neonates, or out-of-scope studies (e.g., different target condition). Lastly, studies focusing on the prediction of septic shock were also excluded as the septic shock was beyond the scope of this review. The extraction was performed by four investigators (MM, BR, MH, and CRJ).

## 2.4. Data Extraction and Synthesis

The following information was extracted from all studies: (i) publication characteristics (first author's last name, publication time), (ii) study design (retrospective, prospective data collection and analysis), (iii) cohort selection (sex, age, prevalence of sepsis), (iv) model selection (machine learning algorithm, platforms, software, packages, and parameters), (v) specifics on the data analyzed (type of data, number of variables), (vi) statistics for model performance (methods to evaluate the model, mean, measure of variance, handling of missing data), and (vii) methods to avoid overfitting as well as any additional external validation strategies. If available, we also reviewed supplementary materials of each study. A full list of extracted variables is provided in **Supplementary Table 2**.

## 2.5. Settings of Prediction Task

Owing to its time sensitivity, setting up the early sepsis prediction task in a clinically meaningful manner is a non-trivial issue. We extracted details on the prediction task as well as the alignment of cases and controls. Given the lack of standardized reporting, the implementation strategies and their reporting vary drastically between studies. Thus, subsequent to gathering all the information, we attempted to create new categories for the sepsis prediction task as well as the case–control alignment. The goal of this new terminology and categories is to increase the comparability between studies.

## 2.6. Assessment of Quality of Reviewed Machine Learning Studies

Based on 14 criteria relevant to the objectives of the review, which we adapted from Qiao (25), the quality of the eligible machine learning studies was assessed. The quality assessment comprised five categories: (1) unmet needs (limits in current machine learning or non-machine learning applications), (2) reproducibility (information on the sepsis prevalence, data and code availability, explanation of sepsis label, feature engineering methods, software/hardware specifications, and hyperparameters), (3) robustness (sample size suited for machine learning applications, valid methods to overcome overfitting, stability of results), (4) generalizability (external data validation), and (5) clinical significance (interpretation of predictors and suggested clinical use; see **Supplementary Table 3**). A quality assessment table was provided by listing "yes" or "no" of corresponding items in each category. MM, BR, MH, and CRJ independently performed the quality assessment. In case of disagreements, ratings were discussed and subsequently, final scores for each publication were determined.

## 2.7. Role of Funding Source

The funding sources of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

# 3. RESULTS

## 3.1. Study Selection

The results of the literature search, including the numbers of studies screened, assessments for eligibility, and articles

[1]This includes peer-reviewed journal articles and peer-reviewed conference proceedings.

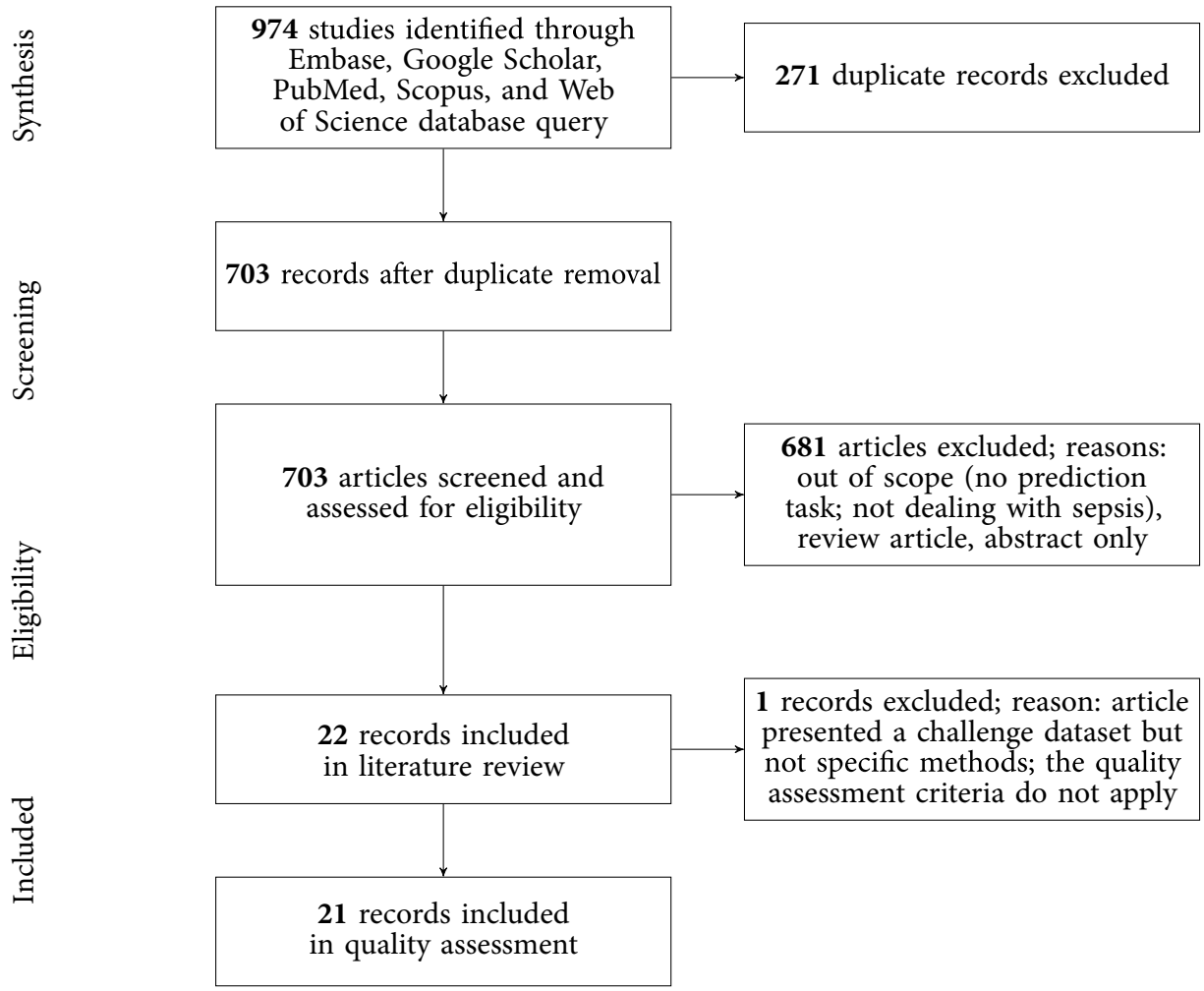



**FIGURE 1 |** PRISMA flowchart of the search strategy. A total of 22 studies were eligible for the literature review and 21 for the quality assessment.

reviewed (with reasons for exclusions at each stage), are presented in **Figure 1**. Out of 974 studies, 22 studies met the inclusion criteria (16–19, 21, 26–42). The majority of excluded studies ($n = 952$) did not meet one or multiple inclusion criteria, such as studying a non-human (e.g., bovine) or a non-adult population (e.g., pediatric or neonatal), focusing on a research topic beyond the current review (e.g., sepsis phenotype identification or mortality prediction), or following a different study design (e.g., case reports, reviews, not-peer reviewed). Detailed information on all included studies are provided in **Table 1**. The inter-rater agreement was excellent ($\kappa = 0.88$).

## 3.2. Study Characteristics

Of the 22 included studies, 21 employed solely retrospective analyses, while one study used both retrospective and prospective analyses (16). Moreover, the most frequent data sources used to develop computational models were MIMIC-II and MIMIC-III ($n = 12$; 54.5%), followed by Emory University Hospital ($n = 5$; 22.7%). In terms of sepsis definition, the majority of the studies employed the Sepsis-2 ($n = 12$; 54.5%) or Sepsis-3 definition ($n = 9$; 40.9%). It is important to note that some studies *modified* the Sepsis-2 or Sepsis-3 definition since all existing definitions have not been intended to specify an exact sepsis onset time (e.g., the employed time window lengths have been varied) (26, 34). In one study (36), sepsis labels were assigned by trained ICU experts. Depending on the definition of sepsis used, and whether subsampling of controls was used to achieve a more balanced class ratio (facilitating the training of machine learning models), the prevalence of patients developing sepsis ranged between 3.3% (See **Table 1**) and 63.6% (**Figure 2**). One study did not report the prevalence (31). Concerning demographics, 9 studies reported the median or mean age, 12 the prevalence of female patients, and solely 1 the ethnicity of the investigated cohorts (**Supplementary Table 4**).

**TABLE 1 |** Overview of included studies.

| | References | Dataset | Sepsis definition | Number of sepsis encounters | Prevalence (%) | Used cohort available | Code for analysis | Code for label | Model | AUROC | Hours before onset | External validation | Data types | Number of variables |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Abromavičius et al. (26) | Emory University Hospital, MIMIC-III | Sepsis-3 (with modified time windows) | 2,932 | 7.3 | Yes | No | No | AdaBoost and Discriminant Subspace Learning | – | – | No | Demographics, labs, vitals | 11 |
| 2 | Barton et al. (17) | MIMIC-III, UCSF | Sepsis-3 | 3,673 | 3.3 | No | No | No | XGBoost | 0.88 | 0 | No | Vitals | 6 |
| 3 | Bloch et al. (27) | RMC | Sepsis-2 related | 300 | 50.0 | No | No | No | Neural Networks, SVM, logistic regression | 0.88 | 4 | No | Vitals | 4 |
| 4 | Calvert et al. (28) | MIMIC-II | Sepsis-2 related | 159 | 11.4 | No | No | No | InSight Algorithm | 0.92 | 3 | No | Demographics, labs, vitals | 9 |
| 5 | Desautels et al. (29) | MIMIC-III | Sepsis-3 | 1,840 | 9.7 | No | No | No | InSight Algorithm | 0.88 | 0 | No | Demographics, vitals | 8 |
| 6 | Futoma et al. (19) | Duke University Health System | Sepsis-2 related | 11,064 | 21.4 | No | No | No | MGP-RNN | 0.91 | 0 | No | Comorbidities, demographics, labs, medications, vitals | 77 |
| 7 | Kaji et al. (18) | MIMIC-III | Sepsis-2 related | 36,176 | 63.6 | Yes | Yes | Yes | LSTM | 0.88 | "Next day" | No | Demographics, labs, medications, vitals | 119 |
| 8 | Kam and Kim (30) | MIMIC-II | Sepsis-2 related | 360 | 6.2 | No | No | No | SepLSTM | 0.99 | 0 | No | Demographics, labs, vitals | 9 |
| 9 | Lauritsen et al. (31) | Danish EHR | Sepsis-2 related | – | – | No | No | No | CNN-LSTM | 0.88 | 0.25 | No | Diagnoses, labs, imaging, medications, vitals, procedures | – |
| 10 | Lukaszewski et al. (32) | Queen Alexandra Hospital | Sepsis-2 related | 25 | 53.2 | No | No | No | MLP | – | – | No | Clinical parameters, cytokine mRNA expression | – |
| 11 | Mao et al. (33) | MIMIC-III, UCSF | Sepsis-2 related | 1,965 | 9.1 | Yes | No | No | InSight Algorithm | 0.92 | 0 | Yes | Vitals | 30 |
| 12 | McCoy and Das (16) | CRMC | Sepsis-3, Severe Sepsis | 407 | 24.4 | No | No | No | InSight Algorithm | 0.91 | – | – | Labs, vitals | – |
| 13 | Moor et al. (21) | MIMIC-III | Sepsis-3 | 570 | 9.2 | Yes | Yes | Yes | MGP-TCN | 0.91 | 0 | No | Labs, vitals | 44 |

*(Continued)*

**TABLE 1** | Continued

| | References | Dataset | Sepsis definition | Number of sepsis encounters | Prevalence (%) | Used cohort available | Code for analysis | Code for label | Model | AUROC | Hours before onset | External validation | Data types | Number of variables |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | Nemati et al. (34) | Emory Healthcare system, MIMIC-III | Sepsis-3 (modified time windows) | 2,375 | 8.6 | No | No | No | Weilbull-Cox proportional hazards model | 0.85 | 4 | Yes | Demographics, vitals | 48 |
| 15 | Reyna et al. (35) | Emory University Hospital, MIMIC-III | Sepsis-3 (modified time windows) | 2,932 | 7.3 | Yes | No | No | – | – | – | Yes | Demographics, labs, vitals | 40 |
| 16 | Schamoni et al. (36) | University Medical Centre Mannheim | Sepsis tag by ICU clinicians | 200 | 32.3 | No | No | No | Non-linear ordinal regression | 0.84 | 4 | No | Comorbidities, demographics, labs, vitals | 55 |
| 17 | Scherpf et al. (37) | MIMIC-III | Sepsis-2 related | 2,724 | 7.7 | No | No | No | RNN-GRU | 0.81 | 3 | No | Labs, vitals | 10 |
| 18 | Shashikumar et al. (38) | Emory Healthcare system | Sepsis-3 | 242 | 22.0 | No | No | No | ElasticNet | 0.78 | 4 | No | Comorbidities, clinical context, demographics, vitals | 17 |
| 19 | Shashikumar et al. (39) | Emory Healthcare system | Sepsis-3 | 100 | 40.0 | No | No | No | SVM | 0.8 | 4 | No | Demographics, comorbidity, clinical context, vitals | 2 |
| 20 | Sheetrit et al. (40) | MIMIC-III | Sepsis-2 related | 1,034 | 41.4 | No | No | No | Temporal Probabilistic Profiles | – | – | No | Demographics, labs, vitals | – |
| 21 | van Wyk et al. (41) | MLH System | Sepsis-2 related | – | 50.0 | No | No | No | Random Forests, RNN | – | – | No | Labs, vitals | 7 |
| 22 | van Wyk et al. (42) | MLH System | Sepsis-2 related | 377 | 50.0 | No | No | No | Random Forests | 0.79 | 0 | No | Vitals | 7 |

*Only if area under the Receiver Operating Characteristic Curve (AUROC) was reported in an early prediction setup, the performance and the corresponding prediction window is reported (in hours before onset). As these windows were highly heterogeneous, to achieve more comparability, we report the minimal hour before onset that was reported. Notably, due to heterogeneous sepsis definition implementations and experimental setups, these metrics likely have low comparability between studies, which is why we deemed a quantitative meta-analysis to be inappropriate.*

*AUROC, area under the ROC curve; CNN-LSTM, convolutional neural network long short-term memory; EHR, electronic health record; ICU, intensive care unit; LSTM, long short-term memory; MGP-RNN, multi-task Gaussian process recurrent neural network; MGP-TCN, multi-task Gaussian process temporal convolutional network; MIMIC, medical information mart for intensive care; MLH, Methodist Le Bonheur Healthcare System; MLP, multilayer perceptron; RMC, Rabin Medical Center; RNN-GRU, recurrent neural net gated recurrent unit; SepLSTM, proper name for LSTM for sepsis; SVM, support vector machine; USCF, University of California San Francisco Health System.*

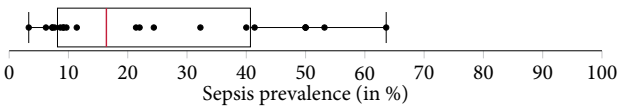



**FIGURE 2 |** A boxplot of the sepsis prevalence distribution of all studies, with the median prevalence being highlighted in red. Note that some studies have subset controls for balancing the class ratios in order to facilitate the training of the machine learning model. Thus, the prevalence in the study cohort (i.e., the subset) can be different from the prevalence of the original data source (e.g., MIMIC-III).

## 3.3. Overview of Machine Learning Algorithms and Data

As shown in **Table 1**, a wide range of predictive models was employed for the early detection of sepsis, with some models being specifically developed for the respective application. Most prominently, various types of neural networks ($n = 9$; 40.9%) were used. This includes recurrent architectures, such as long short-term memory (LSTM) (43) or gated recurrent units (GRU) (44), convolutional networks (45), as well as temporal convolutional networks, featuring causal, dilated convolutions (46, 47). Furthermore, several studies employed boosted tree models ($n = 4$; 18.2%), including XGBoost (48) or random forest (49). As for the data analyzed, the most common data type were vitals ($n = 21$; 95.5%), followed by laboratory values ($n = 13$; 59.1%), demographics ($n = 12$; 54.5%), and comorbidities ($n = 4$; 18.2%). The number of variables included in the respective models ranged between 2 (38) and 119 (18). While reporting the type of variables, four studies failed to report the number of variables included in the models (16, 31, 32, 40).

## 3.4. Model Validation

Approximately 80% of the studies employed one type of cross-validation (e.g., 5-fold, 10-fold, or leave-one-out cross-validation) to avoid overfitting. Additional validation of the models on out-of-distribution ICU data (i.e., external validation) was only performed in three studies (33–35). Specifically, Mao et al. (33) used a dataset provided by the UCSF Medical Center as well as the MIMIC-III dataset to train, validate, and test the *InSight* algorithm. Aiming at developing and validating the *Artificial Intelligence Sepsis Expert (AISE)* algorithm, Nemati et al. (34) created a development cohort using ICU data of over 30,000 patients admitted to two Emory University hospitals. In a subsequent step, the *AISE* algorithm was externally validated on the publicly available MIMIC-III dataset (at the time containing data from over 52,000 ICU stays of more than 38,000 unique patients) (34). Last, the study by Reyna et al. (35) describes the protocol and results of the *PhysioNet/Computing in Cardiology Challenge 2019*. Briefly, the aim of this challenge was to facilitate the development of automated, open-source algorithms for the early detection of sepsis. The PhysioNet/Computing in Cardiology Challenge provided sequestered real-world datasets to the participating researchers for the training, validation, and testing of their models.

## 3.5. Experimental Design Choices for Sepsis Onset Prediction

In this review, we identified two main approaches of implementing sepsis prediction tasks on ICU data. The most-frequent setting ($n = 19$; 86.4%) combines "offline" training with a "horizon" evaluation. Briefly, offline training refers to the fact that the models have access to the entire feature window of patient data. For patients with sepsis, this feature window ranges from hospital admission to sepsis onset, while for the control subjects the endpoint is a matched onset. Alternatively, a prediction window (i.e., a *gap*) between the feature window and the (matched) onset has been employed (27). As for the "horizon" evaluation, the purpose is to determine how early the fitted model would recognize sepsis. To this end, all input data gathered up to $n$ h before onset is provided to the model for the sepsis prediction at a horizon of $n$ h. For studies employing only a single horizon, i.e., predictions preceding sepsis onset by a fixed number of hours, we denote their task as "offline" evaluation in **Table 2**, since there are no sequentially repeated predictions over time. This experimental setup, offline training plus horizon evaluation, is visualized in **Figure 3**. In the second most-frequently used sepsis prediction setting ($n = 2$; 9.1%), both the training and evaluation occur in an "online" fashion. This means that the model is presented with all the data that have been collected until the time point of prediction. The amount of data depends on the spacing of data collection. In order to incentivize early predictions, these timepoint-wise labels can be shifted into the past: in the case of the PhysioNet Challenge dataset, already timepoint-wise labels 6 h before onset are assigned to the positive (sepsis) class (35). For an illustration of an online training and evaluation scenario, refer to **Figure 4**.

Selecting the "onset" for controls (i.e., case–control alignment) is a crucial step in the development of models predicting the onset of sepsis (19). Surprisingly, the majority of the studies ($n = 16$; 72.7%) did not report any details on how the onset matching was performed. For the six studies (27.3%) providing details, we propose the following classification: four employed *random onset matching*, one *absolute onset matching*, and one *relative onset matching* (**Figure 3**, top). As the name indicates, during random onset matching, the onset time of a control is set at a random time of the ICU stay. Often, this time has to satisfy certain additional constraints, such as not being too close to the patient's discharge. The absolute onset matching refers to taking the absolute time since admission until sepsis onset for the case and assigning it as the matched onset time for a control (21). Finally, the relative onset matching is when the matched onset time is defined as the relative time since ICU admission until sepsis onset for the case (50).

## 3.6. Quality of Included Studies

The results of the quality assessment are shown in **Table 3**. One study (35), showcasing the results of the *PhysioNet/Computing in Cardiology Challenge 2019*, was excluded from the quality assessment, which was intended to assess the quality of the implementation and reporting of *specific* prediction models. The quality of the remaining 21 studies ranged from poor (satisfying

**TABLE 2 |** An overview of experimental details: the used sepsis definition, the exact prediction task, and which type of temporal case–control alignment was used (if any).

| | References | Prediction task | Sepsis definition | Case–control alignment | Inclusion criteria |
|---|---|---|---|---|---|
| 1 | Abromavičius et al. (26) | Online training, online evaluation | Sepsis-3 (with modified time windows) | – | – |
| 2 | Barton et al. (17) | Offline training, horizon evaluation | Sepsis-3 | Random onset matching | Inpatients, age ≥18 years, at least one observation per measurement, prediction times between 7 and 2,000 h |
| 3 | Bloch et al. (27) | Offline training, horizon evaluation | Sepsis-2 related: SIRS criteria plus diagnosis of infection | Random onset matching (at least 12 h after admission to the ICU) | age >18 years, admitted to ICU; minimum stay of 12 h in the ICU; patients did not meet SIRS criteria at time of admission to the ICU; Continuous documented measurements were available for at least 12 h for vital signs |
| 4 | Calvert et al. (28) | Offline training, horizon evaluation | Sepsis-2 related: ICD-9 code 995.9 and a 5-h persisting window of fulfilled SIRS | – | Medical ICU, age >18 years, SIRS not fulfilled upon admission, measurements for set of nine variables available |
| 5 | Desautels et al. (29) | Offline training, horizon evaluation, but retrained for each prediction horizon | Sepsis-3 | – | Age ≥15 years, any measurements present, Metavision logging, for cases: sepsis onset between 7 and 500 h after ICU admission, all variables at least once measured, excluded patients that received antibiotics before ICU |
| 6 | Futoma et al. (19) | Offline training, horizon evaluation | Sepsis-2 related: SIRS fulfilled and blood culture drawn and 1 abnormal vital (time windows not stated) | Relative onset matching | Entire EHR cohort included |
| 7 | Kaji et al. (18) | Offline training, horizon evaluation | Sepsis-2 related: SIRS criteria plus ICD-9 code consistent with infection | Fixed length of 14 days in ICU (truncation if longer, zero filling, and masking if shorter) | Individual patient ICU admissions 2 days or longer were identified |
| 8 | Kam and Kim (30) | Offline training, horizon evaluation | Sepsis-2 related: ICD-9 code 995.9 and the first 5-h persisting window of fulfilled SIRS | insufficient detail: during training, 5-h windows are randomly extracted from case before sepsis and entire control stay, during testing it is not stated which data are used for controls | Medical ICU, age >18 years, patient can be checked for 5-h SIRS window plus ICD-9 995.9 code (if only one of the two was available, patients were excluded) |
| 9 | Lauritsen et al. (31) | Offline training, horizon evaluation | Sepsis-2 related: SIRS criteria plus clinically suspected infection | Random onset matching (excluding the first and last 3 h) | Inpatients, admissions ≥3 h, hospital departments with sepsis prevalence ≥2%, ≥1 observations for each vital sign measurement |
| 10 | Lukaszewski et al. (32) | Offline training, offline evaluation (fixed 24-h horizon) | Sepsis-2 related: SIRS criteria plus positive microbiological culture | Insufficient detail (but age-matching between cases and controls; healthy volunteers used as controls) | Blood samples taken daily; last sample on day of diagnosis or last stay in ICU |
| 11 | Mao et al. (33) | Offline training, offline evaluation (single fixed 4-h horizon) | Sepsis-2 related (suspected infection and first hour of fulfilled SIRS criteria), Severe Sepsis: ICD-9 plus SIRS plus organ dysfunction criteria; Septic Shock: ICD-9 plus manually defined conditions | – | Inpatients, age ≥18 years, ≥1 observations for each vital sign measurement, prediction time between 7 and 2,000 h |
| 12 | McCoy and Das (16) | Offline training, evaluation on retrospective dataset, prospective evaluation implemented as risk score | Sepsis-3, Severe Sepsis (SIRS criteria plus 2 organ dysfunction lab values) | – | Age >18 years; two or more sirs criteria during stay (hard to tell "Patient encounters were included in the sepsis-related outcome metrics if they met two or more SIRS criteria at some point during their stay." Is this an inclusion criterion or their label definition?) |
| 13 | Moor et al. (21) | Offline training, horizon evaluation | Sepsis-3 | Absolute onset matching | Age ≥15 years, chart data including ICU admission/discharge time available, Metavision logging, cases: onset at least 7 h into ICU stay |

*(Continued)*

**TABLE 2 |** Continued

| References | Prediction task | Sepsis definition | Case–control alignment | Inclusion criteria |
|---|---|---|---|---|
| 14Nemati et al. (34) | Offline training, horizon evaluation | Sepsis-3 (with modified time windows) | – | Age ≥18 years; sepsis onset not earlier than 4 h within ICU admission |
| 15Reyna et al. (35) | Online training, online evaluation | Sepsis-3 (with modified time windows) | – | ≥8 h of measurements |
| 16Schamoni et al. (36) | Offline training, horizon evaluation as well as prediction of severity (ordinal regression) | Sepsis tag by ICU clinicians via electronic questionnaire | – | Sepsis onset not earlier than on the second day after ICU admission |
| 17Scherpf et al. (37) | Offline training, horizon evaluation | Sepsis-2 related: ICD-9 codes plus SIRS criteria | Random onset matching via drawing fixed size time windows | Age ≥18 years, at least one measurement for SIRS parameters, no sepsis on admission, at least 5 h plus prediction time of measurements |
| 18Shashikumar et al. (38) | Offline training, Offline prediction (single fixed 4-h horizon) | Sepsis-3 | – | – |
| 19Shashikumar et al. (39) | Offline training, Offline prediction (single fixed 4-h horizon) | Sepsis-3 | – | – |
| 20Sheetrit et al. (40) | Offline training, horizon evaluation on two prediction windows (12 and 1 h) | Sepsis-2 related: ICD-9 Codes 995.91 or 995.92 plus antibiotics administered. Onset time is defined as the earliest of either antibiotics prescription or fulfilled qSOFA criteria | Insufficient detail: the paper uses the "equivalent time" as the feature window of the control group | ICU admission, age ≥15 years, for sepsis cases: onset not before third day |
| 21van Wyk et al. (41) | Offline training, horizon evaluation | Sepsis-2 related: SIRS criteria plus suspicion of infection, indicated by the presence of a blood culture and the administration of antibiotics during the encounter, along with relevant ICD10 | Insufficient detail: the paper uses "a given 6-h observational period" for the control group | At least 8 h of continuous data, absence of cardiovascular disease |
| 22van Wyk et al. (42) | Offline training, horizon evaluation | Sepsis-2 related: SIRS criteria plus suspicion of infection, indicated by the presence of a blood culture and the administration of antibiotics during the encounter, along with relevant ICD10 | Insufficient detail: the paper uses "a given 3-h observational period" for the control group | Age >18 years, physiological data available for at least 3 or 6 h, respectively; absence of cardiovascular disease |

*Abbreviations: EHR, electronic health record; ICD-9, International Classification of Disease Version 9; ICU, intensive care unit; qSOFA, quick Sequential Organ Failure Assessment; SIRS, Systemic Inflammatory Response Syndrome.*

≤ 40% of the quality criteria) to very good (satisfying ≥ 90% of the quality criteria). None of the studies fulfilled all 14 criteria. A single criterion was met by 100% of the studies: all studies highlighted the limits in current non-machine-learning approaches in the introduction. Few studies provided the code used for the data cleaning and analysis ($n = 2$; 9.5%), provided data or code for the reproduction of the exact sepsis labels and onset times ($n = 2$; 9.5%), and validated the machine learning models on an external dataset ($n = 3$; 14.3%). For the interpretation, power, and validity of machine learning methods, considerable sample sizes are required. With the exception of one study (32), all studies had sample sizes larger than 50 sepsis patients.

## 4. DISCUSSION

In this study, we systematically reviewed the literature for studies employing machine learning algorithms to facilitate early prediction of sepsis. A total of 22 studies were deemed eligible for the review and 21 were included in the quality assessment. The majority of the studies used data from the MIMIC-III database (13), containing deidentified health data associated with ≈ 60,000 ICU admissions and/or data from Emory University Hospital[2]). With the exception of one, *all* studies used internationally acknowledged guidelines for sepsis definitions, namely Sepsis-2 (51) and Sepsis-3 (1). In terms of the analysis, a wide range of machine learning algorithms were chosen to leverage the patients' digital health data for the prediction of sepsis. Driven by our findings from the reviewed studies, this section first highlights four major challenges that the literature on machine learning driven sepsis prediction is currently facing: (i) asynchronicity, (ii) comparability, (iii) reproducibility, and (iv) circularity. We then discuss the limitations of this study,

[2]The dataset was not publicly available. However, with the 2019 PhysioNet Computing in Cardiology Challenge, a pre-processed dataset from Emory University Hospital has been published (35).

provide some recommendations for forthcoming studies, and conclude with an outlook.

## 4.1. Asynchronicity

While initial studies employing machine learning for the prediction of sepsis have demonstrated promising results (28–30), the literature since has been diverging on which are the most pressing open challenges that need to be addressed to further the goal of early sepsis detection. On the one hand, corporations have been propelling the deployment of the first interventional studies (52, 53), while on the other hand, recent findings have cast doubt on the validity and meaningfulness of the experimental pipeline that is currently being implemented in most retrospective analyses (36). This can be partially attributed to circular prediction settings (for more details, please refer to section 4.4). Ultimately, only the demonstration of favorable outcomes in large prospective randomized controlled trials (RCTs) will pave the way for machine learning models entering the clinical routine. Nevertheless, not every possible choice of model architecture can be tested prospectively due to the restricted sample sizes (and therefore, number of study arms). Rather, the development of these models is generally assumed to occur retrospectively. However, precisely those retrospective studies are facing multiple obstacles, which we are going to discuss next.

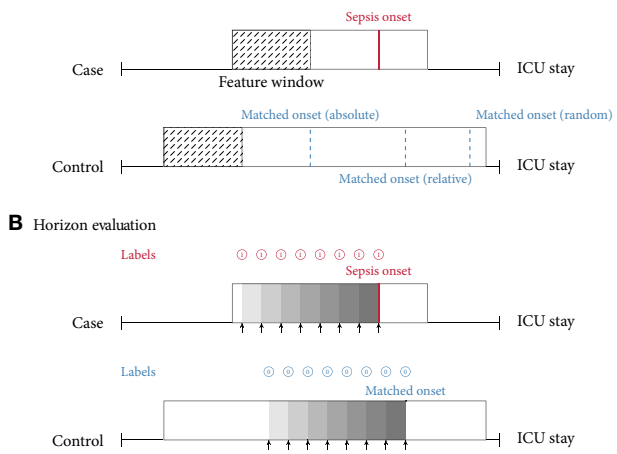



**FIGURE 3 | (A)** Offline training scenario and case–control matching. Every case has a specific sepsis onset. Given a random control, there are multiple ways of determining a matched onset time: (i) *relative* refers to the relative time since intensive care unit (ICU) admission (here, 75% of the ICU stay); (ii) *absolute* refers to the absolute time since ICU admission; (iii) *random* refers to a pseudo-random time during the ICU stay, often with the requirement that the onset is not too close to ICU discharge. **(B)** Horizon evaluation scenario. Given a case and control, with a matched relative sepsis onset, the look-back horizon indicates how *early* a specific model is capable of predicting sepsis. As the (matched) sepsis onset is approached, this task typically becomes progressively easier. Notice the difference in the prediction targets (labels) (shown in red for predicting a case, and blue for predicting a control).

## 4.2. Comparability

Concerning the comparability of the reviewed studies, we note that there are several challenges that have yet to be overcome, namely the choice of (i) prediction task, (ii) case–control onset matching, (iii) sepsis definition, (iv) implementation of a given sepsis definition, and (v) performance measures. We subsequently discuss each of these challenges.

### 4.2.1. Prediction Task

As described in section 3.5, we found that the vast majority of the included papers follow one of two major approaches when implementing the sepsis onset prediction task: Either an offline training step was followed by a horizon evaluation, or both the training and the evaluation were conducted in an online fashion. As one of our core findings, we next highlight the strengths but also the intricacies of these two setups. Considering the most frequently used strategy, i.e., offline training plus horizon evaluation, we found that the horizon evaluation provides valuable information about how early (in hours before sepsis onset) the machine learning model is able to recognize sepsis. However, in order to train such a classifier, the choice of a meaningful time window (and matched onset) for controls is an essential aspect of the study design (for more details, please refer to section 4.2.2). By contrast, the online strategy does not require a matched onset for controls (see **Figure 4**), but it removes the convenience of easily estimating predictive performance for a given prediction horizon (i.e., in hours before sepsis onset). Nevertheless, models trained and evaluated in an online fashion may be more easily deployed in practice, as they are by construction optimized for continuously predicting sepsis while new data arrive. Meanwhile, in the offline setting, the entire classification task is retrospective because all input data are extracted right up until a previously known sepsis onset. Whether a model trained this way would generalize to a prospective setup in terms of predicting sepsis *early* remains to be analyzed in forthcoming studies. In this review, the only study featuring prospective analysis focused on (and improved) prospective targets other than sepsis onset, namely mortality, length of stay, and hospital readmission. Finally, we observed that the online setting also contains a non-obvious design choice, which is absent in the offline/horizon approach: How many hours *before* and *after* a sepsis onset should a positive prediction be considered a true positive or rather a false positive? In other words, how long before or after the onset should a model be incentivized to raise an alarm for sepsis? Reyna et al. (35) proposed a clinical utility score that customizes a clinically motivated reward system for a given positive or negative prediction with respect to a potential sepsis onset. For example, it reflects that late true positive predictions are of little to no clinical importance, whereas late false negatives predictions can indeed be harmful. While such a hand-crafted score may account for a clinician's diagnostic demands, the resulting score remains highly sensitive to the exact specifications for which there is currently neither an internationally accepted standard nor a consensus. Furthermore, in its current form, the proposed clinical utility score is hard to interpret.

### 4.2.2. Case–Control Onset Matching

Futoma et al. (19) observed a drastic drop in performance upon introducing their (relative) case–control onset matching scheme as compared to an earlier version of their study, where the classification scenario compares sepsis onsets with the discharge time of controls (50). Such a matching can be seen as an implicit onset matching, which studies that do not account for this issue tend to default to. This suggests that comparing the data distribution of patients at the time of sepsis onset with the one of controls when being discharged could systematically underestimate the difficulty of the relevant clinical task at hand, i.e., identifying sepsis in an ICU stay. Futoma et al. (19) also remarked that "for non-septic patients, it is not very clinically relevant to include all data up until discharge, and compare predictions about septic encounters shortly before sepsis with predictions about non-septic encounters shortly before discharge. This task would be too easy, as the controls before discharge are likely to be clinically stable." The choice of a matched onset time is therefore crucial and highlights the need for a more uniform reporting procedure of this aspect in the literature. Furthermore, Moor et al. (21) proposed to match the *absolute* sepsis onset time (i.e., perform absolute onset matching) to prevent biases that could arise from systematic differences in the length of stay distribution of sepsis cases and controls (in the worst case, a model could merely re-iterate that one class has shorter stays than the other one, rather than pick up an actual signal in their time series). Finally, **Table 2** lists four studies that employed random onset matching. Given that sepsis onsets are not uniformly distributed over the length the ICU stay (for more details, please refer to section 4.4), this strategy could result in overly distinct data distributions between sepsis cases and non-septic controls.

### 4.2.3. Defining and Implementing Sepsis

A heterogeneous set of existing definitions (and modifications thereof) was implemented in the reviewed studies. The choice of sepsis definition will affect studies in terms of the prevalence of patients with sepsis and the level of difficulty of the prediction task (due to assigning earlier or later sepsis onset times). We note that it remains challenging to fully disentangle all of these factors: on the one side, a larger absolute count of septic patients is expected to be beneficial for training machine learning models (in particular deep neural networks). On the other side, including more patients could make the resulting sepsis cohort a less severe one and harder to distinguish from non-septic ICU patients. Then again, a more inclusive sepsis labeling would result in a higher prevalence (i.e., class balance), which would be beneficial for the training stability of machine learning models. To further illustrate the difficulty of defining sepsis, consider the prediction target *in-hospital mortality*. Even though in-hospital mortality rates (and therefore any subsequent prediction task) vary between cohorts and hospitals, their *definition* typically does not. Sepsis, by contrast, is inherently hard to define, which over the years has led to multiple refinements of clinical criteria (Sepsis 1–3) for trying to capture sepsis in one easy-to-follow, rule-based definition (1, 51, 54). It has been previously shown that applying different sepsis definitions to the same dataset results in largely dissimilar cohorts (55). Furthermore, this specific study found that using Sepsis-3 is too inclusive, resulting in a large cohort showing mild symptoms. By contrast, practitioners have reported that Sepsis-3 is indeed too *restrictive* in that sepsis cannot occur without organ dysfunction any more (55). This suggests that even *within* a specific definition of sepsis, substantial heterogeneity and disagreement in the literature prevails. On top of that, we found that even applying the same definition on the same dataset has resulted in dissimilar cohorts. Most prominently, in **Table 1**, this can be confirmed for studies employing the MIMIC-III dataset. However, the determining factors cannot be easily recovered, as the code for assigning the labels is not available in 19 out of 21 (90.4%) studies employing computer-derived sepsis labels.

Another factor exacerbating comparability is the heterogeneous sepsis prevalence. This is partially influenced by the training setup of a given study, because certain studies prefer balanced datasets for improving the training stability of the machine learning model (27, 41, 42), while others preserve the observed case counts to more realistically reflect how their approach would fare when being deployed in ICU. Furthermore,

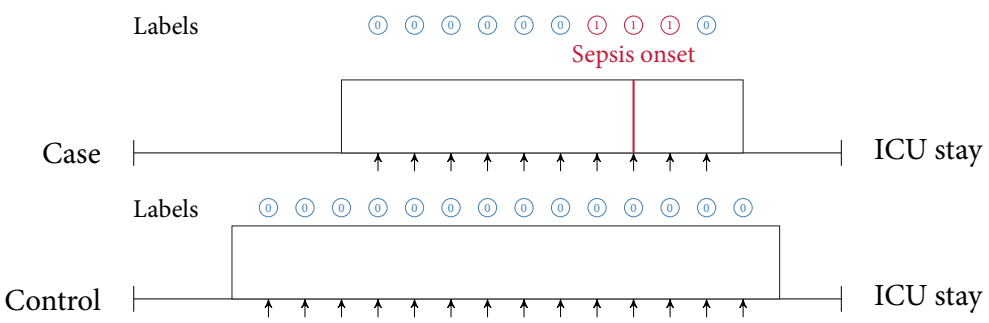



**FIGURE 4 |** Online training and evaluation scenario. Here, the model predicts at regular intervals during an ICU stay (we show predictions in 1-h intervals). For sepsis cases, there is no prima facie notion at which point in time positive predictions ought to be considered as true positive (TP) predictions or false positive (FP) predictions (mutatis mutandis, this applies to negative predictions). For illustrative purposes, here we consider positive predictions up until 1 h before or after sepsis onset (for a case) to be TP.

**TABLE 3 |** Quality assessment of all studies.

| | | Limits of current approaches | Prevalence of sepsis reported | Data availability | Feature engineering methods | Code for analysis | Code for label generation | Platforms/packages reported | Hyperparameters reported | Sample size > 50 | Valid methods to prevent overfitting | Stability of results reported | External data validation | Explanation of predictors | Suggested clinical use | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Abromavičius et al. (26) | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | 50% |
| 2 | Barton et al. (17) | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 57% |
| 3 | Bloch et al. (27) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 71% |
| 4 | Calvert et al. (28) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | 43% |
| 5 | Desautels et al. (29) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | 50% |
| 6 | Futoma et al. (19) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | 50% |
| 7 | Kaji et al. (18) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 93% |
| 8 | Kam and Kim (30) | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | 36% |
| 9 | Lauritsen et al. (31) | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | 57% |
| 10 | Lukaszewski et al. (32) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ | 43% |
| 11 | Mao et al. (33) | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | 64% |
| 12 | McCoy and Das (16) | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | 36% |
| 13 | Moor et al. (21) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 93% |
| 14 | Nemati et al. (34) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | 50% |
| 15 | Schamoni et al. (36) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 57% |
| 16 | Scherpf et al. (37) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | 43% |
| 17 | Shashikumar et al. (38) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | 50% |
| 18 | Shashikumar et al. (39) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | 50% |
| 19 | Sheetrit et al. (40) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | 43% |
| 20 | van Wyk et al. (41) | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | 36% |
| 21 | van Wyk et al. (42) | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | 43% |
| | | 100% | 95% | 19% | 81% | 10% | 10% | 19% | 29% | 95% | 81% | 62% | 14% | 38% | 86% | |
| Study | | Unmet need | Reproducibility | | | | | | | Stability | | | Generalizability | Clinical significance | | Total |

*We excluded Reyna et al. (35) from the assessment because it does presents a dataset challenge rather than a single method, making most of the categories not applicable.*

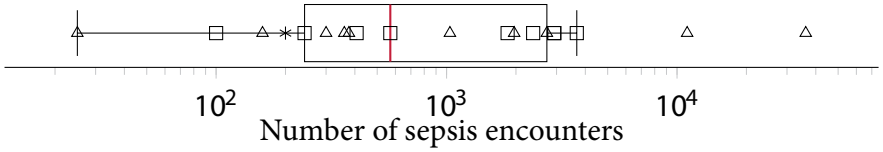



**FIGURE 5 |** A boxplot of the number of sepsis encounters reported by all studies, with the median number of encounters being highlighted in red. Since the numbers feature different orders of magnitude, we employed logarithmic scaling. The marks indicate which definition or modification thereof was used. Sepsis-3: squares, Sepsis-2: triangles, domain expert label: asterisk.

the exact sepsis definition used as well as the applied data pre-processing and filtering steps influence the resulting sepsis case count and therefore the prevalence (21, 55). **Figure 2** depicts a boxplot of the prevalence values of all studies. Of the 22 studies, 10 report prevalences $\leq$ 10%, with the maximum reported prevalence being 63.6% (18). In addition, **Figure 5** depicts the distribution of all sepsis encounters, while also encoding the sepsis definition (or modification thereof) that is being used.

#### 4.2.4. Performance Measures

The last obstacle impeding comparability is the choice of performance measures. This is entangled with the differences in sepsis prevalence: simple metrics, such as accuracy are directly impacted by class prevalence, rendering a comparison of two studies with different prevalence values moot. Some studies report the area under the receiver operating characteristic curve (AUROC, sometimes also reported as AUC). However, AUROC also depends on class prevalence and is known to be less informative if the classes are highly imbalanced (56, 57). The area under the precision–recall curve (AUPRC, sometimes also referred to as average precision) should be reported in such cases, and we observed that $n = 6$ studies already do so. AUPRC is also affected by prevalence but permits a comparison with a random baseline that merely "guesses" the label of a patient. AUROC, by contrast, can be high even for classifiers that fail to properly classify the minority class of sepsis patients. This effect is exacerbated with increasing class imbalance. Recent research suggests reporting the AUPRC of models, in particular in clinical contexts (58), and we endorse this recommendation.

#### 4.2.5. Comparing Studies of Low Comparability

Our findings indicate that quantitatively comparing studies concerned with machine learning for the prediction of sepsis in the ICU is currently a nigh-impossible task. While one would like to perform meta-analyses in these contexts to aggregate an overall trend in performance among state-of-the-art models, at the current stage of the literature this would carry little meaning. Therefore, we currently cannot ascertain the best performing approaches by merely assessing numeric results of performance measures. Rather, we had to resort to *qualitatively* assess study designs in order identify underlying biases, which could lead to overly optimistic results.

### 4.3. Reproducibility

Reproducibility, i.e., the capability of obtaining similar or identical results by independently repeating the experiments described in a study, is the foundation of scientific accountability. In recent years, this foundation has been shaken by the discovery of failures to reproduce prominent studies in several disciplines (59). Machine learning in general is no exception here, and despite the existence of calls to action (60), the field might face a reproducibility crisis (61). The interdisciplinary nature of digital medicine comes with additional challenges for reproducibility (62), foremost of which is the issue of dealing with sensitive data (whereas for many theoretical machine learning papers, benchmark datasets exist), but also the issue of algorithmic details, such as pre-processing. Our quality assessment highlights a lot of potential for improvement here: only two studies (18, 21), both from 2019, share their analysis code and the code for generating a "label" (to distinguish between cases or controls within the scenario of a specific paper). This amounts to $<$ 10% of the eligible studies. In addition, only four studies (18, 21, 26, 33) report results on publicly available datasets (more precisely, the datasets are available for research after accepting their terms and conditions). This finding is surprising, given the existence of high-quality, freely accessible databases, such as MIMIC-III (13) or eICU (63). An encouraging finding of our analysis is that a considerable number of studies ($n = 6$) report hyperparameter details of their models. Hyperparameter refers to any kind of parameter that is model specific, such as the regularization constant and the architecture of a neural network (64). This information is crucial for everyone who attempts to reproduce computational experiments.

### 4.4. Circularity

Considering that the exact sepsis onset is usually unknown, most of the existing works have approximated a plausible sepsis onset via clinical criteria, such as Sepsis-3 (1). However, these criteria comprise a set of rules to apply to vital and laboratory measurements. Schamoni et al. (36) pointed out that using clinical measurements for predicting a sepsis label, which was itself derived from clinical measurements, could potentially be circular (a statistical term referring to the fact that one uses the same data for the selection of a model and its subsequent analysis). This runs the risk being unable to discover unknown aspects of the data, since classifiers may just confirm existing criteria instead of helping to generate new knowledge. In the worst case, a classifier would merely reiterate the guidelines used to define sepsis *without* being able to detect patterns that permit an earlier discovery. To account for this, Schamoni et al. chose a questionnaire-based definition of sepsis and clinical experts

manually labeled the cases and controls. While this strategy may reduce the problem of circularity, a coherent and comprehensive definition of sepsis cannot be easily guaranteed. Notably, Schamoni et al. (36) report very high inter-rater agreement. They assign, however, only *daily* labels, which is in contrast to automated Sepsis-3 labels that are typically extracted in an hourly resolution. Furthermore, it is plausible that even with clinical experts in the loop, some level of (indirect) circularity could still take place, because a clinician would also consult the patients' vital and laboratory measurements in order to assign the sepsis tag, it would merely be less explicit. Since Schamoni et al. (36) proposed a way to circumvent the issue of circularity, this also means that no existing work has empirically assessed the existence (or the relevance) of circularity in machine learning-based sepsis prediction. For Sepsis-3, if the standard 72-h window is used for assessing an increase in SOFA (sequential organ failure assessment score) score, i.e., starting 48 h before suspected infection time until 24 h afterwards, and if the onset happens to occur at the very end of this window, then measurements that go 72 h into the past have influenced this label. Since the SOFA score aggregates the most abnormal measurements of the preceding 24 h (65), Sepsis-3 could even "reach" 96 h into the past. Meanwhile, the distribution of onsets using Sepsis-3 tends to be highly right-skewed, as can be seen in Moor et al. (21), where removing cases with an onset during the first 7 h drastically reduced the resulting cohort size. Therefore, we conjecture that with Sepsis-3, it could be virtually impossible to strictly separate data that are used for assigning the label from data that are used for prediction, without overly reducing the resulting cohort. Finally, the relevance of an ongoing circularity may be challenged given first promising results (in terms of mortality reduction) of the first interventional studies applying machine learning for sepsis prediction prospectively (52), without explicitly accounting for circularity.

## 4.5. Limitations of This Study

A limitation of this review is that our literature search was restricted to articles listed in Embase, Google Scholar, PubMed/Medline, Scopus, and Web of Science. Considering the pace at which the research in this area—in particular, in the context of machine learning—is moving forward, it is likely that the findings of the publications described in this paper will be quickly complemented by further research. The literature search also excluded gray literature (e.g., preprints and reports), the importance of which to this topic is unknown[3], and thus might have introduced another source of search bias. The lack of studies reporting poor performance of machine learning algorithms regarding sepsis onset prediction suggests high probability of publication bias (66, 67). Publication bias is likely to result in studies with more positive results being preferentially submitted and accepted for publication (68). Finally, our review specifically focused on machine learning applications for the prediction of sepsis and severe sepsis. We therefore used a stringent search term that potentially excluded studies pursuing a classical statistical approach of early detection and sepsis prediction.

[3]In the machine learning community, for example, it is common practice to use preprints to disseminate knowledge about novel methods early on.

# 5. RECOMMENDATIONS

This section provides recommendations how to harmonize experimental designs and reporting of machine learning approaches for the early prediction of sepsis in the ICU. This harmonization is necessary to warrant meaningful comparability and reproducibility of different machine learning models, ensure continued model development as opposed to starting from scratch, and establish benchmark models that constitute the state-of-the-art.

As outlined above, only few studies score highly with respect to reproducibility. This is concerning, as reproducibility remains one of the cornerstones of scientific progress (62). The lack of comparability of different studies impedes progress because a priori, it may not be clear which method is suitable for a specific scenario if different studies lack common ground (see also the aforementioned issues preventing a meta-analysis). The way out of this dilemma is to improve reproducibility of a *subset* of a given study. We suggest the following approach: (i) picking an openly available dataset (or a subset thereof) as an additional validation site, (ii) reporting results on this dataset, and (iii) making the code for this analysis available (including models and labels). This suggestion is flexible and still enables authors to showcase their work on their respective private datasets. We suggest that code sharing—within reasonable bounds—should become the *default* for publications as modern machine learning research is increasingly driven by implementations of complex algorithms. Therefore, a prerequisite of being able to replicate the results of any study, or to use it in a comparative setting, is having access to the raw code that was used to perform the experiment. This is crucial, as any pseudocode description of an algorithm permits many different implementations with potentially different runtime behavior and side effects. With only two studies sharing code, method development is stymied. We thus encourage authors to consider sharing their code, for example via platforms, such as GitHub (https://github.com). Even sharing only parts of the code, such as the label generation process, would be helpful in many scenarios and improve comparability. The availability of numerous open source licenses (70) makes it possible to satisfy the constraints of most authors, including companies that want to protect their intellectual property. A recent experiment at the International Conference of Machine Learning (ICML) demonstrated that reviewers and area chairs react favorably to the inclusion of code (71). If code sharing is *not* possible, for example because of commercial interests, there is the option to share binaries, possibly using virtual machines or "containers" (72). Providing containers would satisfy all involved parties: intellectual property rights are retained but additional studies can compare their results.

As for the datasets used in a study, different rules apply. While some authors suggest that peer-reviewed publications should be come with a waiver agreement for open access data (73), we are aware of the complications of sharing clinical data. We think that a reasonable middle ground can be reached by following the suggestion above, i.e., using existing benchmark datasets, such as MIMIC-III (13) to report performance.

**BOX 1** | Recommendations for the practitioner.

| Recommendation | Remarks | Details |
|---|---|---|
| Make code publicly available or usable | A prerequisite of being able to replicate the results of any study, or to use any model in a comparative setting, is having access to the raw code or a binary variant thereof that was used to perform the experiments. Authors are encouraged to share their code, for example via platforms, such as GitHub, or their binaries using container technologies like Docker. | GitHub, Docker |
| Use external validation for the machine learning model | External validation of a classifier is crucial for assessing the model's generalizability. Several publicly available data sources exist that can be used for this purpose. | MIMIC-II, MIMIC-III, eICU, HiRID |
| Provide exact definition of sepsis label | Implementations vary drastically in terms of prevalence and number of sepsis encounters. Thus, reporting the label generation process is essential, particularly when labels deviate from the international definitions of sepsis. For instance, when using the eICU dataset, microbiology measurements are under-reported for defining suspected infection, yet the exact modifications of sepsis implementations have not explicitly been stated (69). | Provide code of how sepsis label was determined. |
| Provide an detailed description of a control and, if applicable, its matched onset | While there is a defined point in time for an event in the sepsis cohorts, it is much more challenging to determine at what time to extract data for a control case when was the non-event. For transparency and replication reasons, it is crucial to provide details on how controls were defined and how the onset was determined. | Provide code of how a control was defined and, if applicable, its matched onset was determined. |
| Make data available | If possible and in compliance with international data protection laws, data sources should be made accessible to bona fide researchers. There are multiple data repositories, which researchers can use to make their data accessible, while complying with data protection laws. | Harvard Dataverse, PhysioNet, Zenodo |
| Ensure comparability of models and their performances | To advance the field, it is important that researchers compare their models to existing models in order to evaluate and compare the performance across different studies. This necessitates improvements in prevalence reporting as well as the choice of different performance metrics. | Report prevalence and AUPRC in addition to other metrics. |
| Use licenses for code | Licenses protect the creators and the users of code. Numerous open source licenses exist, making it possible to satisfy the constraints of most authors, including companies that want to protect their intellectual property. | Apache license, BSD licenses, GPL |

Moreover, we urge authors to report additional details of their experimental setup, specifically the selection of cases and controls and the label generation/calculation process. As outlined above, the case–control matching is crucial as it affects the difficulty (and thus the significance) of the prediction task. We suggest to either follow the absolute onset matching procedure (21), which is simple to implement and prevents biases caused by differences in the length of stay distribution. In any case, forthcoming work should always report their choice of case–control matching. As for the actual prediction task, given the heterogeneous prediction horizons that we observed, we suggest that authors always report performance for a horizon of 3 h or 4 h (in addition to any other performance metrics that are reported). This reporting should always use the AUPRC metric as it is the preferred metric for rare prevalences (74). Last, we want to stress that a description of the inclusion process of patients is essential in order to ensure comparability.

## 6. CONCLUSIONS AND FUTURE DIRECTIONS

This study performed a systematic review of publications discussing the early prediction of sepsis in the ICU by means of machine learning algorithms. Briefly, we found that the majority

of the included papers investigating sepsis onset prediction in the ICU are based on data from the same center, MIMIC-II or MIMIC-III (13), two versions of a high-quality, publicly available critical care database. Despite the data agreement guidelines of MIMIC-III stating that code using MIMIC-III needs to be published (paragraph 9 of the current agreement reads "If I openly disseminate my results, I will also contribute the code used to produce those results to a repository that is open to the research community."), only two studies (18, 21) make their code available. This leaves a lot of room for improvement, which is why we recommend code (or binary) sharing (**Box 1**). Of 22 included studies, only *one* reflects a non-Western (i.e., neither North-American nor European) cohort, pinpointing toward a significant dataset bias in the literature (see **Supplementary Table 4** for an overview of demographical information). In addition to demographic aspects, such as ethnicity, differing diagnostic, and therapeutic policies as well as the availability of input data for prediction are known to impact the generation of the sepsis labels. This challenge hampers additional benchmarking efforts unless more diverse cohorts are included. Moreover, since the prediction task is highly sensitive to minor changes in study specification (including, but not limited to, the sepsis definition and the case–control alignment), the majority of investigated papers do not permit a straightforward reproduction/replication and comparison of their employed cohorts and their presented prediction task. Meta-analyses are therefore impossible, as the reported metrics pertain to different, incomparable scenarios: both prevalence and case counts are highly variable, even on the same dataset, and previous work (19) indicated that minor changes in the experimental setup can substantially affect the difficulty of the prediction task. As a consequence, we are currently not able to identify the most predictive method for recognizing sepsis early, which then ought to be further investigated in prospective trials. All in all, we found this state of the art to leave lots of room for improvement; it would be beneficial to be able to compare different models as to their generalizability, in particular when deploying machine learning algorithms in a prospective study. We see our paper as a "call to arms" for the community and hope that our recommendations are taken in the spirit of improving this task together.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://github.com/BorgwardtLab/sepsis-prediction-review.

## AUTHOR CONTRIBUTIONS

MM, BR, and CJ contributed substantially to the data acquisition, extraction, analysis (i.e., quality assessment), and interpretation. Furthermore, they drafted the review article. MH made a substantial contributions to data interpretation (i.e., quality assessment) and participated in revising the review article critically for important intellectual content. KB made a significant contributions to the study conception and revised the review article critically for important intellectual content. All authors contributed to the article and approved the submitted version.

## FUNDING

This project was supported by the Strategic Focal Area Personalized Health and Related Technologies (PHRT) of the ETH Domain for the SPHN/PHRT Driver Project Personalized Swiss Sepsis Study (Borgwardt, #2017-110) and the Swiss National Science Foundation (Ambizione Grant, PZ00P3-186101, Jutzeler). Moreover, this work was funded in part by the Alfried Krupp Prize for Young University Teachers of the Alfried Krupp von Bohlen und Halbach-Stiftung (KB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## ACKNOWLEDGMENTS

This manuscript has been released as a preprint at *medRxiv* (75).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.607952/full#supplementary-material

## REFERENCES

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*. (2016) 315:801–10. doi: 10.1001/jama.2016.0287
2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet*. (2020) 395:200–11. doi: 10.1016/S0140-6736(19)32989-7
3. Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock 2012. *Crit Care Med*. (2013) 41:580–637. doi: 10.1097/CCM.0b013e31827e83af
4. Kaukonen KM, Bailey M, Suzuki S, Pilcher D, Bellomo R. Mortality related to severe sepsis and septic shock among critically ill patients in Australia and New Zealand, 2000–2012. *J Am Med Assoc*. (2014) 311:1308–16. doi: 10.1001/jama.2014.2637
5. Hotchkiss RS, Moldawer LL, Opal SM, Reinhart K, Turnbull IR, Vincent JL. Sepsis and septic shock. *Nat Rev Dis Primers*. (2016) 2:16045. doi: 10.1038/nrdp.2016.45
6. Ferrer R, Martin-Loeches I, Phillips G, Osborn TM, Townsend S, Dellinger RP, et al. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Crit Care Med*. (2014) 42:1749–55. doi: 10.1097/CCM.0000000000000330
7. Weiss SL, Fitzgerald JC, Balamuth F, Alpern ER, Lavelle J, Chilutti M, et al. Delayed antimicrobial therapy increases mortality and organ

dysfunction duration in pediatric sepsis. *Crit Care Med*. (2014) 42:2409. doi: 10.1097/CCM.0000000000000509
8. Pruinelli L, Westra BL, Yadav P, Hoff A, Steinbach M, Kumar V, et al. Delay within the 3-hour surviving sepsis campaign guideline on mortality for patients with severe sepsis and septic shock. *Crit Care Med*. (2018) 46:500. doi: 10.1097/CCM.0000000000002949
9. Lever A, Mackenzie I. Sepsis: definition, epidemiology, and diagnosis. *BMJ*. (2007) 335:879–883. doi: 10.1136/bmj.39346.495880.AE
10. Al Jalbout N, Troncoso Jr R, Evans JD, Rothman RE, Hinson JS. Biomarkers and molecular diagnostics for early detection and targeted management of sepsis and septic shock in the emergency department. *J Appl Lab Med*. (2019) 3:724–9. doi: 10.1373/jalm.2018.027425
11. Parlato M, Philippart F, Rouquette A, Moucadel V, Puchois V, Blein S, et al. Circulating biomarkers may be unable to detect infection at the early phase of sepsis in ICU patients: the CAPTAIN prospective multicenter cohort study. *Intensive Care Med*. (2018) 44:1061–70. doi: 10.1007/s00134-018-5228-3
12. Faix JD. Biomarkers of sepsis. *Crit Rev Clin Lab Sci*. (2013) 50:23–36. doi: 10.3109/10408363.2013.764490
13. Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. (2016). doi: 10.1038/sdata.2016.35
14. Fleuren LM, Klausch TL, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med*. (2020) 46:383–400. doi: 10.1007/s00134-019-05872-y
15. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health*. (2020) 2:179–91. doi: 10.1016/S2589-7500(20)30018-2
16. McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual*. (2017) 6:e000158. doi: 10.1136/bmjoq-2017-000158
17. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, et al. Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Comput Biol Med*. (2019) 109:79–84. doi: 10.1016/j.compbiomed.2019.04.027
18. Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, Costa AB, et al. An attention based deep learning model of clinical events in the intensive care unit. *PLoS ONE*. (2019) 14:e0211057. doi: 10.1371/journal.pone.0211057
19. Futoma J, Hariharan S, Heller K, Sendak M, Brajer N, Clement M, et al. *An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. Vol. 68 of Proceedings of Machine Learning Research*. PMLR (2017). p. 243–54.
20. Li SCX, Marlin BM. A scalable end-to-end gaussian process adapter for irregularly sampled time series classification. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R. editors. *Advances in Neural Information Processing Systems*. Curran Associates, Inc., (2016). 29, p. 1804–12.
21. Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K. *Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping. Vol. 106 of Proceedings of Machine Learning Research*. PMLR (2019). p. 2–26.
22. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. (2015) 4:1. doi: 10.1186/2046-4053-4-1
23. Harzing A. *Publish or Perish Software*. (2007). Available online at: https://harzingcom/resources/publish-or-perish
24. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med*. (2005) 37:360–3.
25. Qiao N. A systematic review on machine learning in sellar region diseases: quality and reporting items. *Endocr Connect*. (2019) 8:952–60. doi: 10.1530/EC-19-0156
26. Abromavičius V, Plonis D, Tarasevičius D, Serackis A. Two-stage monitoring of patients in intensive care unit for sepsis prediction using non-overfitted machine learning models. *Electronics*. (2020) 9:1133. doi: 10.3390/electronics9071133
27. Bloch E, Rotem T, Cohen J, Singer P, Aperstein Y. Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction. *J Healthc Eng*. (2019) 2019:5930379. doi: 10.1155/2019/5930379
28. Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. *Comput Biol Med*. (2016) 74:69–73. doi: 10.1016/j.compbiomed.2016.05.003
29. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform*. (2016) 4:e28. doi: 10.2196/medinform.5909
30. Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. *Comput Biol Med*. (2017) 89:248–55. doi: 10.1016/j.compbiomed.2017.08.015
31. Lauritsen SM, Kalør ME, Kongsgaard EL, Lauritsen KM, Jørgensen MJ, Lange J, et al. Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artif Intell Med*. (2020) 104:101820. doi: 10.1016/j.artmed.2020.101820
32. Lukaszewski RA, Yates AM, Jackson MC, Swingler K, Scherer JM, Simpson A, et al. Presymptomatic prediction of sepsis in intensive care unit patients. *Clin Vacc Immunol*. (2008) 15:1089–94. doi: 10.1128/CVI.00486-07
33. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open*. (2018) 8:e017833. doi: 10.1136/bmjopen-2017-017833
34. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. (2018) 46:547. doi: 10.1097/CCM.0000000000002936
35. Reyna MA, Josef C, Seyedi S, Jeter R, Shashikumar SP, Westover MB, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Crit Care Med*. (2019). 48:210–217. doi: 10.1097/CCM.0000000000004145
36. Schamoni S, Lindner HA, Schneider-Lindner V, Thiel M, Riezler S. Leveraging implicit expert knowledge for non-circular machine learning in sepsis prediction. *Artif Intell Med*. (2019) 100:101725. doi: 10.1016/j.artmed.2019.101725
37. Scherpf M, Gräßer F, Malberg H, Zaunseder S. Predicting sepsis with a recurrent neural network using the MIMIC III database. *Comput Biol Med*. (2019) 113:103395. doi: 10.1016/j.compbiomed.2019.103395
38. Shashikumar SP, Li Q, Clifford GD, Nemati S. Multiscale network representation of physiological time series for early prediction of sepsis. *Physiol Meas*. (2017) 38:2235. doi: 10.1088/1361-6579/aa9772
39. Shashikumar SP, Stanley MD, Sadiq I, Li Q, Holder A, Clifford GD, et al. Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. *J Electrocardiol*. (2017) 50:739–43. doi: 10.1016/j.jelectrocard.2017.08.013
40. Sheetrit E, Nissim N, Klimov D, Shahar Y. Temporal probabilistic profiles for sepsis prediction in the ICU. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY (2019). p. 2961–9. doi: 10.1145/3292500.3330747
41. Van Wyk F, Khojandi A, Kamaleswaran R. Improving prediction performance using hierarchical analysis of real-time data: a sepsis case study. *IEEE J Biomed Health Inform*. (2019) 23:978–86. doi: 10.1109/JBHI.2019.2894570
42. van Wyk F, Khojandi A, Mohammed A, Begoli E, Davis RL, Kamaleswaran R. A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier. *Int J Med Inform*. (2019) 122:55–62. doi: 10.1016/j.ijmedinf.2018.12.002
43. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735
44. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv*. (2014) 14061078. doi: 10.3115/v1/D14-1179
45. Fukushima K, Miyake S, Ito T. Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Trans Syst Man Cybernet*. (1983) 36:826–34. doi: 10.1109/TSMC.1983.6313076

46. Lea C, Flynn MD, Vidal R, Reiter A, Hager GD. Temporal convolutional networks for action segmentation and detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017). p. 156–65. doi: 10.1109/CVPR.2017.113
47. Oord Avd, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. Wavenet: a generative model for raw audio. *arXiv*. (2016) 160903499.
48. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY (2016). p. 785–94. doi: 10.1145/2939672.2939785
49. Kam HT. Random decision forest. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition*. (1995). p. 278–82.
50. Futoma J, Hariharan S, Heller K. Learning to detect sepsis with a multitask Gaussian process RNN classifier. *arXiv*. (2017) 170604152.
51. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, et al. 2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference. *Intensive Care Med*. (2003) 29:530–8. doi: 10.1007/s00134-003-1662-x
52. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res*. (2017) 4:234. doi: 10.1136/bmjresp-2017-000234
53. Burdick H, Pino E, Gabel-Comeau D, McCoy A, Gu C, Roberts J, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. *BMJ Health Care Inform*. (2020) 27:e100109. doi: 10.1136/bmjhci-2019-100109
54. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, et al. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*. (1992) 101:1644–55. doi: 10.1378/chest.101.6.1644
55. Johnson AE, Aboab J, Raffa JD, Pollard TJ, Deliberato RO, Celi LA, et al. A comparative analysis of sepsis identification methods in an electronic database. *Crit Care Med*. (2018) 46:494. doi: 10.1097/CCM.0000000000002965
56. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr*. (2008) 17:145–51. doi: 10.1111/j.1466-8238.2007.00358.x
57. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. (2015) 10:e0118432. doi: 10.1371/journal.pone.0118432
58. Pinker E. Reporting accuracy of rare event classifiers. *NPJ Digit Med*. (2018) 1:56. doi: 10.1038/s41746-018-0062-0
59. Baker M. 1,500 Scientists lift the lid on reproducibility. *Nature*. (2016) 533:452–4. doi: 10.1038/533452a
60. Crick T, Hall BA, Ishtiaq S. "Can i implement your algorithm?" A model for reproducible research software. *arXiv*. (2014) 1407.5981.
61. Hutson M. Artificial intelligence faces reproducibility crisis. *Science*. (2018) 359:725–6. doi: 10.1126/science.359.6377.725
62. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. *NPJ Digit Med*. (2019) 2:2. doi: 10.1038/s41746-019-0079-z
63. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data*. (2018) 5:180178. doi: 10.1038/sdata.2018.178
64. Wu J, Chen XY, Zhang H, Xiong LD, Lei H, Deng SH. Hyperparameter optimization for machine learning models based on Bayesian optimization. *J Electron Sci Technol*. (2019) 17:26–40. doi: 10.11989/JEST.1674-862X.80904120
65. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. *The SOFA (Sepsis-related Organ Failure Assessment) Score to Describe Organ Dysfunction/Failure*. Heidelberg: Springer (1996). doi: 10.1007/s001340050156
66. Dickersin K, Chalmers I. Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the WHO. *J R Soc Med*. (2011) 104:532–8. doi: 10.1258/jrsm.2011.11k042
67. Kirkham JJ, Altman DG, Williamson PR. Bias due to changes in specified outcomes during the systematic review process. *PLoS ONE*. (2010) 5:e9810. doi: 10.1371/journal.pone.0009810
68. Joober R, Schmitz N, Annable L, Boksa P. Publication bias: what are the challenges and can they be overcome? *J Psychiatry Neurosci*. (2012) 37:149. doi: 10.1503/jpn.120065
69. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. (2018) 24:1716–20. doi: 10.1038/s41591-018-0213-5
70. Rosen L. *Open Source Licensing: Software Freedom and Intellectual Property Law*. Upper Saddle River, NJ: Prentice Hall (2004).
71. Chaudhuri K, Salakhutdinov R. *The ICML 2019 Code-at-Submit-Time Experiment*. (2019). Available online at: https://medium.com/@kamalika_19878/the-icml-2019-code-at-submit-time-experiment-f73872c23c55
72. Elmenreich W, Moll P, Theuermann S, Lux M. Making computer science results reproducible–a case study using Gradle and Docker. *PeerJ*. (2018) 6:e27082v1. doi: 10.7287/peerj.preprints.27082v1
73. Hrynaszkiewicz I, Cockerill MJ. Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. *BMC Res Notes*. (2012) 5:494. doi: 10.1186/1756-0500-5-494
74. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol*. (2015) 68:855–9. doi: 10.1016/j.jclinepi.2015.02.010
75. Moor M, Rieck B, Horn M, Jutzeler C, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. *medRxiv*. (2020). doi: 10.1101/2020.08.31.20185207


**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2021 Moor, Rieck, Horn, Jutzeler and Borgwardt. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Public Health

BRIEF RESEARCH REPORT
published: 16 June 2021
doi: 10.3389/fpubh.2021.662842

# Data Sharing in Southeast Asia During the First Wave of the COVID-19 Pandemic

*Arianna Maever L. Amit*[1,2,3*†], *Veincent Christian F. Pepito*[2†], *Bernardo Gutierrez*[4,5†] *and Thomas Rawson*[4*†]

[1] Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States, [2] School of Medicine and Public Health, Ateneo de Manila University, Pasig, Philippines, [3] College of Medicine, University of the Philippines Manila, Manila, Philippines, [4] Department of Zoology, University of Oxford, Oxford, United Kingdom, [5] School of Biological and Environmental Sciences, Universidad San Francisco de Quito USFQ, Quito, Ecuador

OPEN ACCESS

**Edited by:**
Adrian Egli,
University Hospital of Basel, Switzerland

**Reviewed by:**
Mario Coccia,
National Research Council (CNR), Italy
Wesley Cota,
Universidade Federal de Viçosa, Brazil

***Correspondence:**
Arianna Maever L. Amit
alamit@up.edu.ph
Thomas Rawson
t.rawson@imperial.ac.uk

**†ORCID:**
Arianna Maever L. Amit
orcid.org/0000-0003-4571-400X
Veincent Christian F. Pepito
orcid.org/0000-0001-5391-3784
Bernardo Gutierrez
orcid.org/0000-0002-9220-2739
Thomas Rawson
orcid.org/0000-0001-8182-4279

**Specialty section:**
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

**Received:** 01 February 2021
**Accepted:** 11 May 2021
**Published:** 16 June 2021

**Citation:**
Amit AML, Pepito VCF, Gutierrez B and Rawson T (2021) Data Sharing in Southeast Asia During the First Wave of the COVID-19 Pandemic. Front. Public Health 9:662842. doi: 10.3389/fpubh.2021.662842

**Background:** When a new pathogen emerges, consistent case reporting is critical for public health surveillance. Tracking cases geographically and over time is key for understanding the spread of an infectious disease and effectively designing interventions to contain and mitigate an epidemic. In this paper we describe the reporting systems on COVID-19 in Southeast Asia during the first wave in 2020, and highlight the impact of specific reporting methods.

**Methods:** We reviewed key epidemiological variables from various sources including a regionally comprehensive dataset, national trackers, dashboards, and case bulletins for 11 countries during the first wave of the epidemic in Southeast Asia. We recorded timelines of shifts in epidemiological reporting systems and described the differences in how epidemiological data are reported across countries and timepoints.

**Results:** Our findings suggest that countries in Southeast Asia generally reported precise and detailed epidemiological data during the first wave of the pandemic. Changes in reporting rarely occurred for demographic data, while reporting shifts for geographic and temporal data were frequent. Most countries provided COVID-19 individual-level data daily using HTML and PDF, necessitating scraping and extraction before data could be used in analyses.

**Conclusion:** Our study highlights the importance of more nuanced analyses of COVID-19 epidemiological data within and across countries because of the frequent shifts in reporting. As governments continue to respond to impacts on health and the economy, data sharing also needs to be prioritised given its foundational role in policymaking, and in the implementation and evaluation of interventions.

**Keywords: COVID- 19, epidemiological data, Southeast Asia, emerging infectious disease, data sharing**

## INTRODUCTION

In December 2019, an outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was reported in Wuhan, China and was determined to cause the novel coronavirus disease 2019 (COVID-19). The World Health Organization (WHO) declared the outbreak to be a Public Health Emergency of International Concern on 30 January 2020, and subsequently a pandemic on 11 March 2020.

The impact of the pandemic required robust research to understand the novel virus and develop effective mitigation and containment strategies (1–3). In February, the WHO in collaboration with the Global Research Collaboration for Infectious Disease Preparedness and Response (GLOPID-R) developed the Global Research Roadmap in response to the pandemic and identified priority research areas (4). These included: (a) product development for improvement of clinical processes; (b) shedding, natural history of disease; (c) monitoring of phenotypic change and adaptation; (d) immunity; and (e) disease models (4). Since then, a vast number of research has been produced on the clinical aspects of the disease, non-pharmaceutical interventions (NPIs), and public health (3). There has also been interest in the role of the environment (5–10), use of machine learning techniques and digital technologies (11–16), and government and policy responses (17–23).

To effectively respond to public health emergencies, there is a need for timely and accurate reporting of statistics and data sharing as highlighted in the recent Ebola and Zika epidemics (24–27). To this end, the Principles for Data Sharing in Public Health Emergencies consisting of timeliness, ethics, equitability, accessibility, transparency, fairness, and quality have been developed and introduced (26, 28, 29). The Global Research Roadmap also identified data sharing as a cross-cutting research priority that spans all other key topics (4). As of writing, however, the current evidence into the quality and availability of data is severely limited, with studies focusing primarily around descriptions of data sources or comments on the importance of data and data sharing (24, 30–42). One research group has examined the data availability for 507 COVID-19 patients reported in January, finding that the majority of information was provided by social media and news outlets (43). Other than this example, there is no other original work that investigates the issues surrounding data availability and data sharing practices during the pandemic. In Southeast Asia, only one study on data sharing during disease outbreaks has been carried out (44). The study evaluated data quality and timeliness of outbreak reporting in Cambodia, Lao PDR, Myanmar, and Vietnam for dengue, food poisoning and diarrhea, severe diarrhea, diphtheria, measles, H5N1 influenza, H1N1 influenza, rabies, and pertussis. Further, it highlighted the broad differences observed in the data quality and timeliness between participating countries, concluding that any international data-curating attempts must be versatile enough to accommodate these.

Ongoing research into the epidemiology of SARS-CoV-2 depends entirely on access to regularly updated and factor-rich data. The benefits and importance of data sharing practices have been well-documented during previous outbreaks. In the ongoing COVID-19 crisis, government organisations, public health agencies, and research groups are responding to the call for rapid data sharing by providing data and curating detailed real-time databases that are readily and publicly accessible (30–32). Data from various groups have informed more than 100,000 papers on COVID-19 (45). Despite progress in reporting and sharing data, the scale of the global pandemic presents its own unique challenges. First, there are ethical and privacy considerations that need to be balanced carefully against the potential impact of open data sharing. Second, there is a clear lack of capacity and often appropriate computational infrastructure that may make data sharing in real time unfeasible and burdensome (26, 27). Such challenges may result in changes in the quality and detail of data reporting between and within countries over time as their respective health systems become increasingly overwhelmed (33). The majority of countries are now routinely reporting the number of confirmed cases and deaths attributed to COVID-19, with the country-wide cumulative totals readily accessible from databases such as the one curated by Johns Hopkins University (30). However, the breadth of further information reported by each country is less understood. Access to demographic and geographic information of cases in particular is critically important in the context of informing policy response, as these provide greater insights into how subgroups of the population in different areas are affected by the disease. Understanding how and when these data are provided is critical to ensuring that modelling efforts and government response are well-informed. Further, understanding global responses to the COVID-19 pandemic will be of increasing relevance as countries begin to develop updated post-pandemic disease response frameworks. Being able to compare and contrast how different countries responded and provided information in the early stages of the pandemic will be crucial in designing better response and reporting pipelines for future global health crises.

Our work thus aimed to explore the scale of data reporting across the broader pandemic timeline by describing the ways in which various countries in a geographic region report COVID-19 data and how the detail of data reporting changed over time. We reviewed detailed epidemiological data from Southeast Asian countries and tracked how countries' reporting of COVID-19 data has shifted. We further evaluated differences in reporting between countries and described the accessibility of epidemiological data during the first wave in 2020. By providing these types of information, researchers may be able to conduct better and more nuanced analyses of epidemiological data of COVID-19. Further, our research provides wider insight into the data pipeline from government to researchers, and how it has adapted over time. This timeline provides greater context to the specific findings of subsequent data-driven research, highlighting areas and time periods where particular data feeds are likely to be particularly biased or data-sparse. We are also able to recommend, based upon our findings, prioritising the use of the early-case histories of specific countries for the calculation of demographic-specific disease parameters. By highlighting particular regions where specific data are available, such as travel history, hospitalisation times and symptom-tracking, we are also able to identify ideal further

topics of research in the ongoing attempts to fight the spread of COVID-19.

# METHODS

## Study Design

We conducted an observational study to describe and track changes in reporting of epidemiological data during the COVID-19 pandemic in 11 countries in Southeast Asia, namely Brunei, Cambodia, Indonesia, Lao PDR, Malaysia, Myanmar, Philippines, Singapore, Thailand, Timor-Leste, and Vietnam. Such a design allows us to compare the data reporting practices between different countries through time as the pandemic progresses (46).

## Data Sources and Compilation

We focused on reporting mechanisms of individual level COVID-19 data from the aforementioned 11 countries in Southeast Asia. The region is characterised by archipelagos and comprises more than 8.0% of the world's population. During the first wave of the pandemic, these 11 countries contributed about 1.3% of the cases to the global count of more than 2.3 million cases on April 20.

We initially reviewed the data of the Open COVID-19 Data Curation Group's centralised repository containing individual-level information on patients with laboratory-confirmed COVID-19 (47). These included data on the following variables deemed essential in monitoring pandemics: (a) Key dates, which include the date of travel, date of onset of symptoms, date of confirmation of infection, date of admission to hospital, and date of outcome; (b) Demographic information including the age and sex of cases; (c) Geographic information on domicile and travel history at the highest resolution available down to the district level; (d) Any additional information such as symptoms and 'contact tracing data' (i.e., a record of exposure to infected individuals) (47). The collection of data on these variables mirrors the minimum data to be collected for a line list of pandemic influenza cases obtained from surveillance systems, as suggested by the WHO (48). Other sources, such as the interactive dashboard by Johns Hopkins University (30), do not provide detailed individual-level information and hence were not used in this study. At the time of the conduct of this study, the said centralised repository was manually maintained by a number of individuals, and therefore would have potentially missed some information about the COVID-19 positive individuals, particularly occupation that was not recorded in the repository. To validate and augment the data from the centralised repository, we reviewed other relevant and official data sources of each country in different formats including: government trackers and dashboards that report close to real-time data, downloadable PDF reports, downloadable CSV files, and official social media accounts of governmental or public health institutions (**Supplementary Table 1**). In addition, we reviewed data from news agencies, pre-prints, and peer-reviewed research articles that contained information on COVID-19 cases in the country. We reviewed all possible publicly available data sources from the date when the first confirmed case was reported in the country, and up to April 20. We only collected data at one timepoint, on April 20, and therefore could only use information available then. No updates on the reporting of key epidemiological variables were made for this study.

## Data Interpretation and Analysis

We documented trends and changes in how key epidemiological variables were reported by 11 Southeast Asian countries throughout the study period from January 23 to April 20. The reporting methodologies of each country could broadly be separated into three distinct time periods, defined by specific milestones in each country's data reporting. The first time period or "first reporting of cases" (T0) for all countries was the date at which the country reported its first COVID-19 case. Following this, the "first change in reporting" (T1) was the time when the information format was changed from the first report based on available data during the study period. This was primarily characterised by countries establishing a formal channel by which to declare subsequent confirmed cases of COVID-19, as opposed to (T0), where cases were primarily reported via news reports and/or government briefings. Any further changes in the level of detail, also referred in this paper as granularity for geographic data and precision for both demographic and temporal data, in the reporting of any of the epidemiological variables were considered as a "change in reporting" and were noted as a subsequent time period (**Supplementary Table 2**). This was characterised by countries further updating and altering their previously established formal case declaration channel as their respective data pipelines changed. The "last observed change in reporting" (T2) was the last documented change up to April 20. We also noted the number of cases in each timepoint. In this paper, we only present results on the "first reporting of cases" (T0), "first observed change in reporting" (T1), and "last observed change in reporting" (T2).

We then explored the differences in reporting of demographic, geographic, and temporal data across countries at three key timepoints: at the time they first reported cases (T0), at the time when the reporting first changed (T1), and at the last observed change in reporting (T2). Any change in the level of granularity or precision in reporting is noted. We present these differences for each epidemiological variable classified into: (a) demographic data; (b) geographic data; and (c) temporal data. Data for other epidemiological variables are presented in **Supplementary Figures 1–5**. We present in **Supplementary Figure 6** a summary of what information each country had for each timepoint (T0, T1, T2).

# RESULTS

## Shifts in Reporting of Epidemiological Data During the First Wave

The first Southeast Asian country to report a COVID-19 case was Thailand on January 23. Singapore, Malaysia, Cambodia, Vietnam and the Philippines subsequently reported cases on or before the WHO declared COVID-19 a Public Health Emergency of International Concern (PHEIC) on January 30. Indonesia,

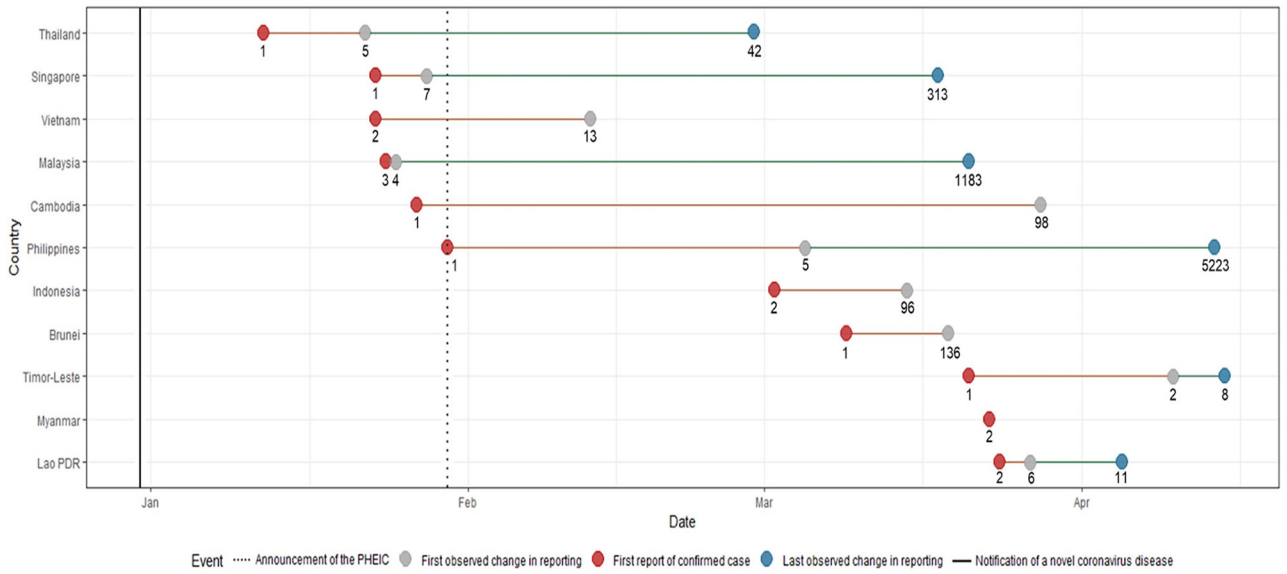



**FIGURE 1 |** Timeline of key events with corresponding number of cases during the first wave of the COVID-19 pandemic. The notification of a novel coronavirus disease in late December and the announcement of the PHEIC are denoted by lines. Shifts in reporting are defined by a change in level of detail and precision in any of the epidemiological variables (**Supplementary Table 2**). Each key reporting shift is denoted by a colored circle.

Brunei, Timor-Leste, Myanmar and Lao PDR reported their first cases of COVID-19 in March (**Figure 1**).

Malaysia had the shortest time between reporting of the first case and first change in reporting of epidemiological data. Only a day after their first reported case, more detailed reports on the occurrence of symptoms, and dates of symptom onset and hospitalisation were provided. Similar improvements in terms of the level of granularity and precision in reporting data were also noted for the following countries: Philippines eventually reported comorbidities for some patients, Singapore and Vietnam eventually reported data on occupation, and Timor-Leste eventually reported travel history data. As case numbers increased, several countries provided less detailed information. By March 15, when 96 cases had been identified, Indonesia ceased reporting individual-level data and switched to aggregate data (i.e., number of cases per day). Timor-Leste followed by April 15, when it had 8 recorded cases. The first and the last changes in reporting were the same for Indonesia and Brunei, while Myanmar was the only country that consistently reported individual-level COVID-19 epidemiologic data since reporting its first two cases on March 23 until April 20.

## Differences in the Granularity and Precision of Reporting Across Countries

There were minimal changes in the reporting of demographic data among countries. The majority of countries reported age and sex except for Timor-Leste, and only Indonesia shifted from a more precise reporting of age and sex to less detailed reporting (**Figures 2A,B**). We observed more changes in the reporting of occupation (**Figure 2C**); Indonesia only provided occupation data at the time of reporting of first cases, while Singapore and Vietnam included data on occupation of COVID-19 patients at later timepoints.

Location information on domicile and travel history differed across countries and timepoints. While all 11 countries provided domicile information (**Figure 3A**), only Singapore provided precise-level addresses. Both Indonesia and Malaysia initially provided city-level information and shifted to less granular reporting. For Indonesia, province-level data was being reported by March 15 when it reached 96 cases. Meanwhile, Malaysia started reporting province-level data on March 21 when it reached 1,183 confirmed cases. On the other hand, the information coming from some countries initially presented less granularity or lower geographic resolution: Lao PDR initially reported country-level information, Thailand initially reported province-level addresses and Vietnam initially reported city-level addresses; eventually all three countries reported precise address data. There were less differences observed for travel location data reporting across countries, but also more shifts observed over time (**Figure 3B**). Most (8 of 11) provided city-level information of the travel history; only Myanmar provided country-level information, while both Indonesia and Timor-Leste provided no information at the time of reporting their first cases. Only Timor-Leste shifted to a more granular level of reporting over time, while Brunei, Cambodia, Malaysia, Philippines, Singapore and Thailand reported less granular data. Lao PDR shifted reporting travel histories from city-level information when it reported its first two cases to no information being shared when it had six confirmed cases, and then to country-level travel history data when it had reported 11 cases.

For all temporal variables, countries reported either precise dates or no dates at all. At the start of each country's first case,

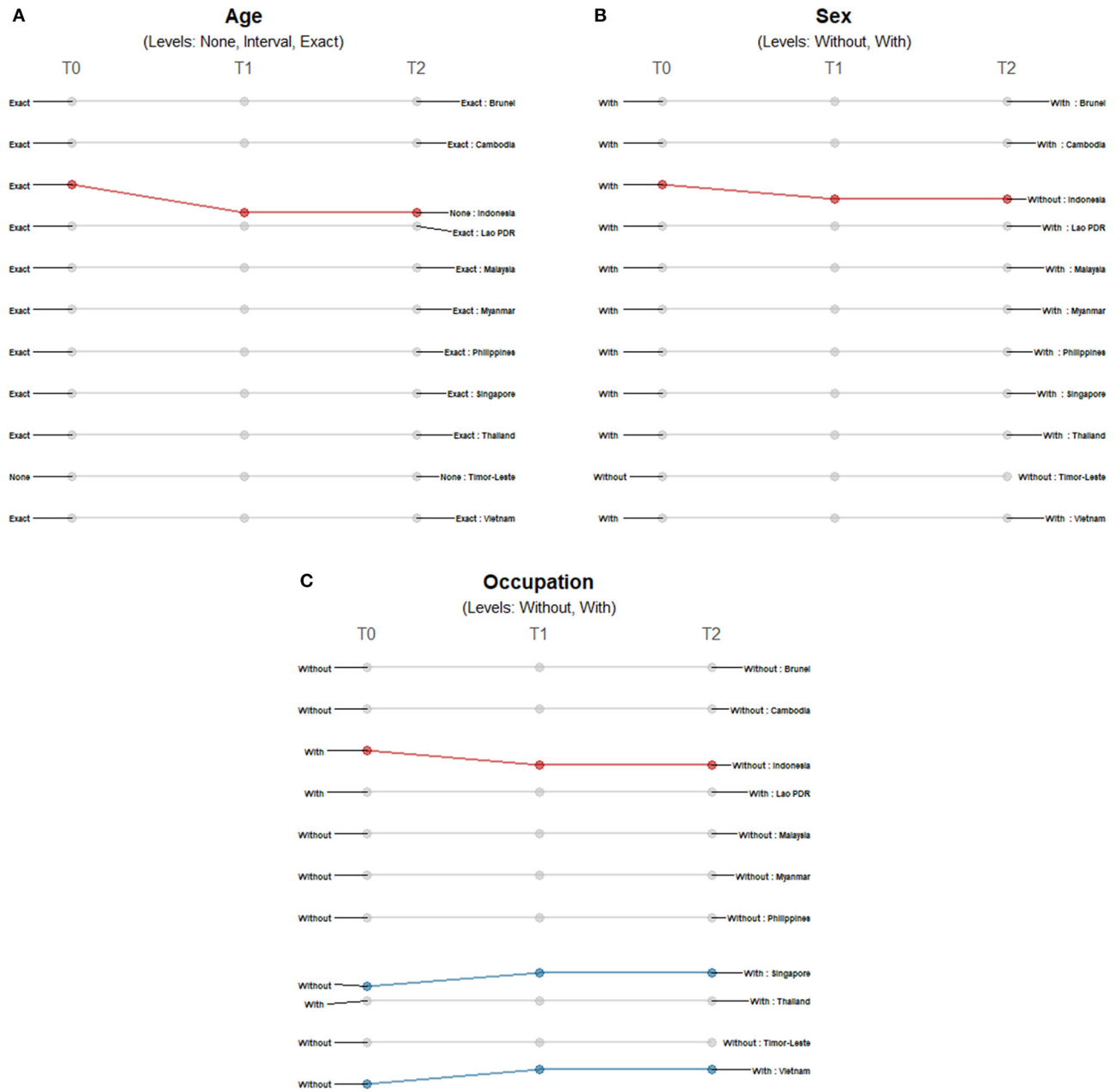



**FIGURE 2 |** Differences in the level of precision in reporting demographic data: **(A)** age, **(B)** sex, and **(C)** occupation over three timepoints. Only those countries with changes in the level of detail and precision of reporting are highlighted. Each country may shift reporting at any timepoint: at the first reporting of cases' (T0), "first observed change in reporting" (T1), and "last observed change in reporting" (T2). Each country may report less precise data indicated by a decreasing slope (red) or more precise data indicated by an increasing slope (blue) consistently over time. Reporting may not be consistent across timepoints with shifts between different levels of precision (yellow) or reporting may not have changed at all during the study period (gray). The levels of precision are indicated for each epidemiological variable. Age has three levels while both sex and occupation are binary variables.

the majority of countries provided travel history dates except for Brunei, Indonesia, and Timor-Leste (**Figure 4A**). Only Brunei shifted to reporting dates for the succeeding timepoints while Malaysia, Philippines, and Singapore stopped reporting dates as cases increased. Lao PDR repeatedly shifted between reporting travel dates and excluding this information. The precision of reporting symptom onset dates also varied across countries and timepoints (**Figure 4B**). Cambodia, Indonesia and Timor-Leste never reported such information, while Brunei, Myanmar, and Vietnam consistently reported specific dates when symptoms

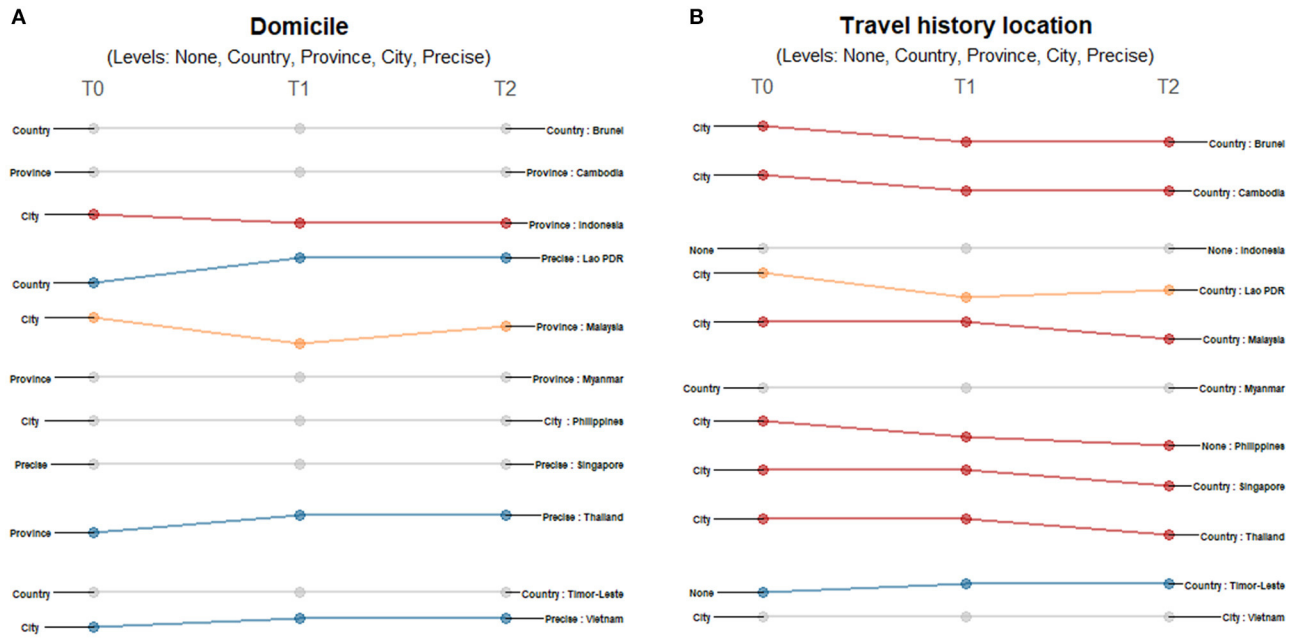



**FIGURE 3 |** Differences in the level of granularity in reporting geographic data: **(A)** domicile, and **(B)** travel history location over three timepoints. Only those countries with changes in the level of granularity or geographic resolution of reporting are highlighted. Each country may shift reporting at any timepoint: at the first reporting of cases' (T0), "first observed change in reporting" (T1), and "last observed change in reporting" (T2). Each country may report less granular data indicated by a decreasing slope (red) or more granular data indicated by an increasing slope (blue) consistently over time. Reporting may not be consistent across timepoints with shifts between different levels of granularity (yellow) or reporting may not have changed at all during the study period (gray). The levels of granularity are indicated for each epidemiological variable. All geographic data have five levels of granularity/geographic resolution: none, country, province, city, and precise.

presented. Malaysia provided day information in the succeeding timepoints while Philippines, Singapore, and Thailand eventually stopped reporting the date of symptom onset. Lao PDR repeatedly shifted between reporting of dates to no reporting. Date of confirmation showed consistent reporting in all countries except Thailand, which stopped its reporting when it had 42 cases (**Figure 4C**). Several countries initially reported the date of admission except for Brunei, Cambodia, Indonesia, Malaysia, and Timor-Leste (**Figure 4D**). Only Thailand had a shift in reporting dates of discharge, recovery, or death - reporting this information only in late February when it had 42 cases (**Figure 4E**).

## DISCUSSION

Responding to calls for data sharing and transparency, most governments in Southeast Asia established publicly available sources of COVID-19 individual-level information. This commitment to data sharing and reporting allowed the comparison of the different data reporting practices of the countries in the region. We found that countries in Southeast Asia have different reporting practices since the start of the pandemic and during the first months of its progression. Overall, reporting of epidemiological data in Southeast Asia is precise and detailed. Many variables were consistently maintained throughout the initial outbreak period, but those with changes in reporting started early with case counts as low as four to as high as 136. There was little to no change in reporting of demographic data while changes in reporting of geographic and temporal variables were frequent and unpredictable as the pandemic progressed. Further, we find that changes in the level of precision in reporting does not only depend on case numbers, but also on the policies and interventions implemented. Comparisons across countries for different epidemiologic variables showed that national governments may shift to a less or more precise reporting of data as dictated by the burden of COVID-19 in the communities and/or their national response. As an example, Indonesia started reporting aggregate data less than two weeks after their first case was reported. Their government did not implement a nationwide lockdown, but rather focused on scaling up capacity, treating patients and supporting economic recovery. Conversely, Lao PDR, Thailand and Vietnam reported more precise demographic and geographic data at the end of the study period compared to how they reported their first cases. The national governments of these countries established mechanisms to quickly identify and isolate cases and their contacts requiring detailed contact tracing data. Our findings also show that most countries reported more precise information towards the end of the study period, but some variables such as travel history location were reported with less detail compared to the increased granularity for domicile data. These trends in travel history data highlight the shift in priorities of the governments in the

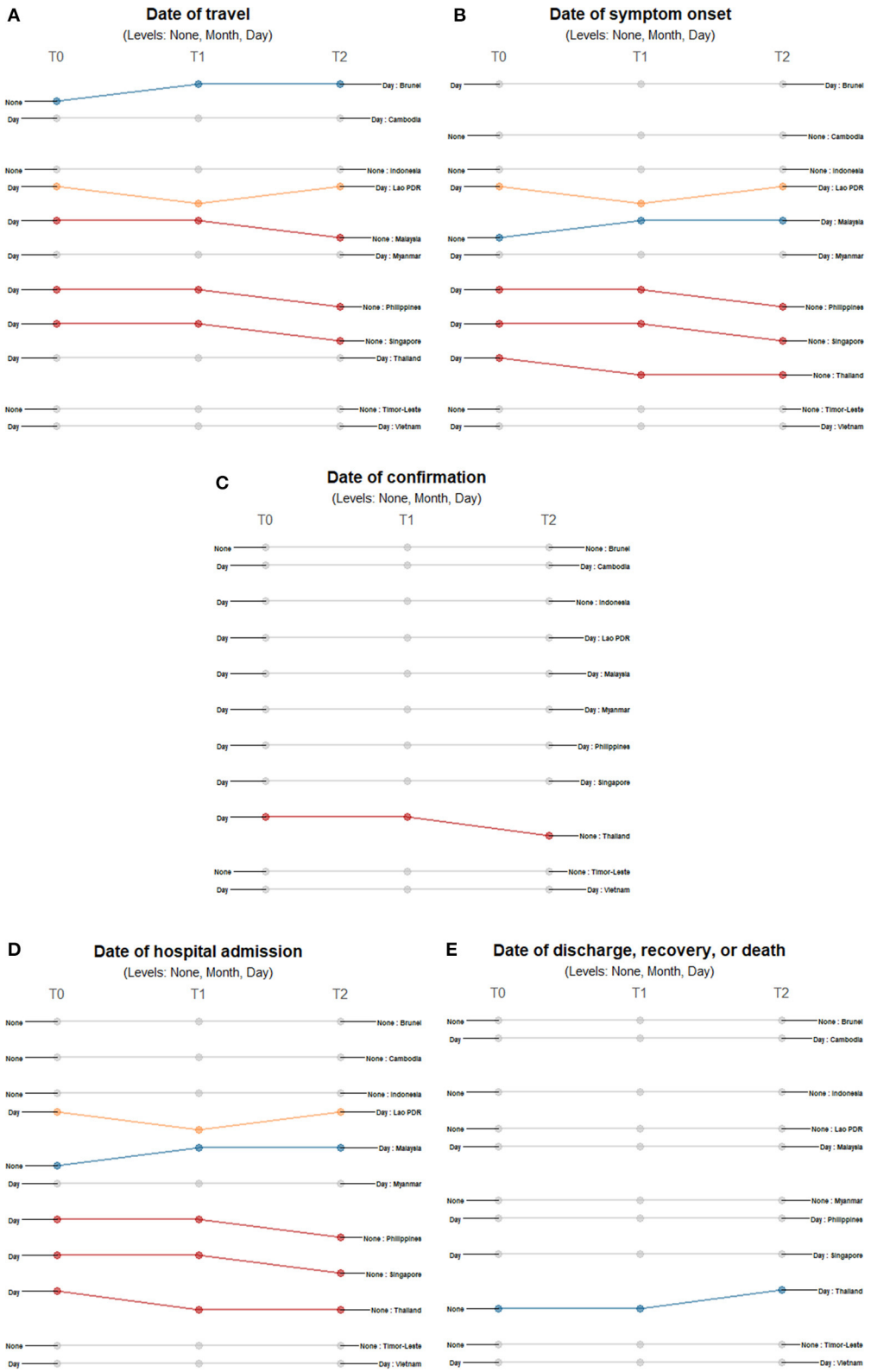



**FIGURE 4 |** Differences in the level of precision in reporting temporal data: **(A)** date of travel, **(B)** date of symptom onset, **(C)** date of confirmation, **(D)** date of hospital admission, and **(E)** date of outcome over three timepoints. Only those countries with changes in the level of detail and precision of reporting are highlighted.

*(Continued)*

**FIGURE 4 |** Each country may shift reporting at any timepoint: at the first reporting of cases' (T0), "first observed change in reporting" (T1), and "last observed change in reporting" (T2). Each country may report less precise data indicated by a decreasing slope (red) or more precise data indicated by an increasing slope (blue) consistently over time. Reporting may not be consistent across timepoints with shifts between different levels of precision (yellow) or reporting may not have changed at all during the study period (gray). The levels of precision are indicated for each epidemiological variable. All date variables have three levels of precision: none, month, and day.

region towards managing local transmission. Southeast Asian countries implemented travel restrictions early, therefore having fewer imported cases and less need for precise travel history data (49).

Data on dates of symptom onset, confirmation, admission, and outcome (discharge, recovery, or death) are important in estimating disease burden and forecasting health service needs. Dates of confirmation and outcome (discharge, recovery, or death) were reported consistently by most countries. This reflects the effective system of governments to register all confirmed patients in their database upon entry and exit in the healthcare system. However, we found that dates of symptom onset and hospital admission were no longer reported at the end of the observation period for some countries. The reporting of less precise dates could be attributed to the increasing incidence of COVID-19, which could have overwhelmed data reporting mechanisms of the countries, particularly because individual patient follow-up requires symptom onset dates to be accurately logged. Governments thus need to establish systems that allow accurate and fast reporting of detailed temporal data. Lack of precision could adversely affect the quality of mathematical models and other analyses, which are used to forecast demand for health services and make decisions. This consequently impacts the responses to COVID-19 at a national and subnational level, which is of greater concern among low- and middle-countries (LMICs) that already have fragile health systems. Our findings provide insights on how different health systems respond to the pandemic. Consequently, these could be used to guide how publicly available data are analysed, used, and interpreted.

Most countries reported COVID-19 data daily, with unclear reporting frequencies only being observed for Brunei, Lao PDR, and Timor-Leste. These countries do not report new cases every day because of the low number of new daily cases leading to days where no additional cases are confirmed. As they only provide updates on days when new COVID-19 cases are confirmed, their frequency of providing data updates on COVID-19 is thus irregular. Countries primarily reported individual-level data in either HTML and PDF formats, which necessitates scraping and extraction before such data could be used in analyses. During the study period, only Thailand provided a downloadable CSV format of their data. Ready-to-use data formats are important as these allow the public and scientific community to rapidly view and analyse country-specific information.

Shifts in reporting, especially from a detailed level of reporting to aggregated data, provide a challenge for accurately comparing epidemiological situations between countries, more so for understanding disease dynamics and guiding government actions. In China, it has been shown that changes in reporting have impacted modelling results of the transmission parameters of COVID-19 (45). Further, as the pandemic progresses and epidemiological information becomes increasingly less available, analyses of detailed case counts that cover the entire duration of the epidemic may not be feasible (32). In Nigeria, a forecasting algorithm has been proposed for use in policy responses given the limited data and constrained data infrastructures in the country (50). In Spain where data have been aggregated as early as May, there have been challenges in conducting age-specific time series, understanding disease transmission, and recommending interventions and policies (42). These three examples are evidence that detailed COVID-19 data are necessary, not only for research purposes, but to ultimately guide policies that avert cases and deaths in the country.

An important limitation of this study is the collection of data at only one timepoint in April. This may not accurately reflect the daily reporting situation of the 11 Southeast Asian countries when the pandemic started. Another limitation is the absence of any assessment on data quality. This evaluation was not carried out because of the fast progression of the pandemic with corresponding rapid changes in data reporting. The lack of an up-to-date and complete line list also prevents a thorough assessment of data quality. Lastly and most importantly, an evaluation of data quality also requires the consideration of other indicators such as flexibility, representativeness, data security and system stability to provide a more accurate picture of health systems and disease surveillance systems (44). These, information are not readily available and require more resources to be collected. Despite such caveats, however, this study is the first to systematically describe and compare reporting of important epidemiological data for COVID-19 across countries during the first wave. Our findings will allow researchers to conduct more nuanced analyses using epidemiological data of COVID-19.

## CONCLUSION

Reporting systems in the region have been quickly established and countries provided detailed individual-level data during the first wave. This pandemic highlights the critical role of timely, accurate, and precise data sharing during outbreaks of global scale. Some concerns regarding data sharing remain, such as data privacy and public criticisms (26, 27). Given that sharing of data is needed for evidence-informed policies and interventions, maintaining and strengthening data reporting systems should still be a priority of countries (51–53). For the purposes of surveillance on emerging infectious diseases, we recommend that

governments coordinate data collection and reporting so that data are as comparable as possible between countries. Countries may also benefit from reporting data in a fully open access format that is readily available and in machine-readable formats to accommodate new epidemics and context-specific information. Hopefully, more governments will come to share precise data to allow more nuanced analyses. This will provide an opportunity to better understand the disease and how best to respond to the pandemic.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analysed in this study. Data can be found here: https://github.com/beoutbreakprepared/nCoV2019.

## AUTHOR CONTRIBUTIONS

AA and VP wrote the initial draft of the manuscript with inputs from BG and TR. All authors contributed to the study concept and design, data compilation, analysis, interpretation, and critically revised the report and approved the final version for submission.

## FUNDING

The work was supported through an Engineering and Physical Sciences Research Council (EPSRC) (https://epsrc.ukri.org/) Systems Biology studentship award (EP/G03706X/1) to TR. This project was also supported in part by the Oxford Martin School. The funders had no role in study design, data collection and analysis, and decision to publish or preparation of the manuscript.

## ACKNOWLEDGEMENTS

We thank Moritz U.G. Kraemer and the Open COVID-19 Data Working Group for their support and insights. The full list of curators and contributors making up the Open COVID-19 Data Working Group is provided at: https://github.com/beoutbreakprepared/nCoV2019.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.662842/full#supplementary-material

## REFERENCES

1. Fauci AS, Lane HC, Redfield RR. Covid-19 — Navigating the uncharted. *N Engl J Med.* (2020) 2:1268–9. doi: 10.1056/NEJMe2002387
2. Lipsitch M, Swerdlow DL, Finelli L. Defining the epidemiology of Covid-19 — studies needed. *N Engl J Med.* (2020) 3:1194–6;. doi: 10.1056/NEJMp2002125
3. Doanvo A, Qian X, Ramjee D, Piontkivska H, Desai A, Majumder M. Machine learning maps research needs in COVID-19 literature. *Patterns.* (2020) 1:100123. doi: 10.1016/j.patter.2020.100123
4. World Health Organization. *A Coordinated Global Research Roadmap*. (2020) Available online at: https://www.who.int/blueprint/priority-diseases/key-action/Roadmap-version-FINAL-for-WEB.pdf?ua=1 (accessed April 22, 2021).
5. Phillips CA. Compound climate risks in the COVID-19 pandemic. *Nat Clim Chang.* (2020) 10:3. doi: 10.1038/s41558-020-0804-2
6. Coccia M. Effects of the spread of COVID-19 on public health of polluted cities: results of the first wave for explaining the dejà vu in the second wave of COVID-19 pandemic and epidemics of future vital agents. *Environ Sci Pollut Res.* (2021) 28:19147–54. doi: 10.1007/s11356-020-11662-7
7. Coccia M. An index to quantify environmental risk of exposure to future epidemics of the COVID-19 and similar viral agents: theory and practice. *Environ Res.* (2020) 191:110155. doi: 10.1016/j.envres.2020.110155
8. Merow C, Urban MC. Seasonality and uncertainty in global COVID-19 growth rates. *Proc Natl Acad Sci USA.* (2020) 117:27456–64. doi: 10.1073/pnas.2008590117
9. Liu X, Huang J, Li C, Zhao Y, Wang D, Huang Z, et al. The role of seasonality in the spread of COVID-19 pandemic. *Environ Res.* (2021) 195:110874. doi: 10.1016/j.envres.2021.110874
10. Ching J, Kajino M. Rethinking air quality and climate change after COVID-19. *Int J Environ.* (2020) 17:5167. doi: 10.3390/ijerph17145167
11. Gomes D, Serra G. Machine learning model for computational tracking and forecasting the COVID-19 dynamic propagation. *IEEE J Biomed Health Inform.* (2021) 25:8. doi: 10.1109/JBHI.2021.3052134
12. Tyagi I. COVID-19: journey so far and deep insight using crowdsourced data in India. *MAPAN-J Metrol Soc I.* (2021) 14:33–46. doi: 10.1007/s12647-020-00416-y
13. Tu B, Wei L, Jia Y, Qian J. Using Baidu search values to monitor and predict the confirmed cases of COVID-19 in China: evidence from Baidu index. *BMC Infect Dis.* (2021) 21:98. doi: 10.1186/s12879-020-05740-x
14. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. *Nat Med.* (2020) 26:1183–92. doi: 10.1038/s41591-020-1011-4
15. de Souza TA, de Almeida Medeiros A, Barbosa IR, de Vasconcelos Torres G. Digital technologies for monitoring infected people, identifying contacts and tracking transmission chains in the corona virus disease 2019 pandemic: a protocol for a systematic review. *Medicine.* (2020) 99:e23744. doi: 10.1097/MD.0000000000023744
16. Khataee H, Scheuring I, Czirok A, Neufeld Z. Effects of social distancing on the spreading of COVID-19 inferred from mobile phone data. *Sci Rep.* (2021) 9:1–8. doi: 10.1038/s41598-021-81308-2
17. Coccia M. The relation between length of lockdown, numbers of infected people and deaths of Covid-19, and economic growth of countries: lessons learned to cope with future pandemics similar to Covid-19 and to constrain the deterioration of economic system. *Sci Total Environ.* (2021) 775:145801. doi: 10.1016/j.scitotenv.2021.145801
18. Han E, Tan MMJ, Turk E, Sridhar D, Leung GM, Shibuya K, et al. Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe. *Lancet.* (2020) 396:1525–34. doi: 10.1016/S0140-6736(20)32007-9
19. Summers DJ, Cheng DH-Y, Lin PH-H, Barnard DLT, Kvalsvig DA, Wilson PN, et al. Potential lessons from the Taiwan and New Zealand health responses to the COVID-19 pandemic. *Lancet Reg Health West Pac.* (2020) 4:100044. doi: 10.1016/j.lanwpc.2020.100044
20. Oh J, Lee J-K, Schwarz D, Ratcliffe HL, Markuns JF, Hirschhorn LR. National response to COVID-19 in the republic of Korea and lessons learned for other countries. *Health Syst Reform.* (2020) 6:e1753464. doi: 10.1080/23288604.2020.1753464
21. Xuan Tran B, Thi Nguyen H, Quang Pham H, Thi Le H, Thu Vu G, Latkin CA, et al. Capacity of local authority and community on epidemic response in Vietnam: implication for COVID-19 preparedness. *Saf Sci.* (2020) 130:104867. doi: 10.1016/j.ssci.2020.104867

22. de Villiers C, Cerbone D, Van Zijl W. The South African government's response to COVID-19. *J Public Budg.* (2020) 32:797–811. doi: 10.1108/JPBAFM-07-2020-0120
23. Amit AML, Pepito VCF, Dayrit MM. Early response to COVID-19 in the Philippines. *Western Pac Surveill Response J.* (2021) 12:5. doi: 10.5365/wpsar.2020.11.1.014
24. Dye C, Bartolomeos K, Moorthy V, Kieny MP. Data sharing in public health emergencies: a call to researchers. *Bull World Health Organ.* (2016) 94:158. doi: 10.2471/BLT.16.170860
25. Pearce N, Vandenbroucke JP, VanderWeele TJ, Greenland S. Accurate statistics on COVID-19 are essential for policy guidance and decisions. *Am J Public Health.* (2020) 110:949–51. doi: 10.2105/AJPH.2020.305708
26. Littler K, Boon W-M, Carson G, Depoortere E, Mathewson S, Mietchen D, et al. Progress in promoting data sharing in public health emergencies. *Bull World Health Organ.* (2017) 95:243. doi: 10.2471/BLT.17.192096
27. Chretien J-P, Rivers CM, Johansson MA. Make data sharing routine to prepare for public health emergencies. *PLoS Med.* (2016) 13:e1002109. doi: 10.1371/journal.pmed.1002109
28. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data.* (2016) 3:160018. doi: 10.1038/sdata.2016.18
29. Global Research Collaboration for Infectious Disease Preparedness (GLOPID-R) Data Sharing Working Group. *Principles For Data Sharing in Public Health Emergencies.* (2018). Available online at: https://wellcome.figshare.com/articles/Principles_for_Data_Sharing_in_Public_Health_Emergencies/4733590 (accessed May 7, 2020).
30. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect.* (2020) 20:533–4. doi: 10.1016/S1473-3099(20)30120-1
31. Xu B, Gutierrez B, Mekaru S, Sewalk K, Goodwin L, Loskill A, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci Data.* (2020) 7:106. doi: 10.1038/s41597-020-0448-0
32. Chen E, Lerman K, Ferrara E. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. *JMIR Public Health Surveill.* (2020) 6:e19273. doi: 10.2196/19273
33. Moorthy V, Henao Restrepo AM, Preziosi M-P, Swaminathan S. Data sharing for novel coronavirus (COVID-19). *Bull World Health Organ.* (2020) 98:150. doi: 10.2471/BLT.20.251561
34. Fegan G, Cheah PY. Solutions to COVID-19 data sharing. *Lancet Digit Health.* (2021) 3:e6. doi: 10.1016/S2589-7500(20)30273-9
35. Gardner L, Ratcliff J, Dong E, Katz A. A need for open public data standards and sharing in light of COVID-19. *Lancet Infect.* (2021) 21:e80. doi: 10.1016/S1473-3099(20)30635-6
36. Galvin CJ, Fernandez-Luque L, Li Y-C. Accelerating the global response against the exponentially growing COVID-19 outbreak through decent data sharing. *Diagn.* (2020). doi: 10.1016/j.diagmicrobio.2020.115070. [Epub ahead of print].
37. Cosgriff CV. Data sharing in the era of COVID-19. *Lancet Digit Health.* (2020) 2:1. doi: 10.1016/S2589-7500(20)30082-0
38. Paul S, Chatterjee MK. Data sharing solutions for biobanks for the COVID-19 pandemic. *Biopreserv Biobank.* (2020) 18:581–6. doi: 10.1089/bio.2020.0040
39. Curioso WH, Carrasco-Escobar G. Collaboration in times of COVID-19: the urgent need for open-data sharing in Latin America. *BMJ Health Care Inform.* (2020) 27:e100159. doi: 10.1136/bmjhci-2020-100159
40. Foraker RE, Lai AM, Kannampallil TG, Woeltje KF, Trolard AM, Payne PRO. Transmission dynamics: data sharing in the COVID-19 era. *Learn Health Sys.* (2021) 5. doi: 10.1002/lrh2.10235. [Epub ahead of print].
41. Aguiar ERGR, Navas J, Pacheco LGC. The COVID-19 diagnostic technology landscape: efficient data sharing drives diagnostic development. *Front Public Health.* (2020) 8:309. doi: 10.3389/fpubh.2020.00309
42. Trias-Llimós S, Alustiza A, Prats C, Tobias A, Riffe T. The need for detailed COVID-19 data in Spain. *Lancet Public Health.* (2020) 5:e576. doi: 10.1016/S2468-2667(20)30234-6
43. Sun K, Chen J, Viboud C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *Lancet Digit Health.* (2020) 2:e201–8. doi: 10.1016/S2589-7500(20)30026-1
44. Lawpoolsri S, Kaewkungwal J, Khamsiriwatchara A, Sovann L, Sreng B, Phommasack B, et al. Data quality and timeliness of outbreak reporting system among countries in greater mekong subregion: challenges for international data sharing. *PLoS Negl Trop Dis.* (2018) 12:e0006425. doi: 10.1371/journal.pntd.0006425
45. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Burdick D, et al. *CORD-19: The COVID-19 Open Research Dataset.* (2020) Available online at: http://arxiv.org/abs/2004.10706 (accessed October 21, 2020).
46. Rothman K, Greenland S, Lash T. *Modern Epidemiology*. 3th ed. St. Philadelphia, PA: Lippincott Williams and Wilkins (2008).
47. Xu B, Kraemer MUG, Xu B, Gutierrez B, Mekaru S, Sewalk K, et al. Open access epidemiological data from the COVID-19 outbreak. *Lancet Infect.* (2020) 20:534. doi: 10.1016/S1473-3099(20)30119-5
48. World Health Organization. *WHO guidance for surveillance during an influenza pandemic.* (2017). Available online at: https://www.who.int/influenza/preparedness/pandemic/WHO_Guidance_for_surveillance_during_an_influenza_pandemic_082017.pdf (accessed May 16, 2020).
49. World Health Organization. *Pandemic Influenza Preparedness and Response: A WHO Guidance Document.* (2009) Available online at: https://www.ncbi.nlm.nih.gov/books/NBK143061/ (accessed August 20, 2020).
50. Abdulmajeed K, Adeleke M, Popoola L. Online forecasting of COVID-19 cases in Nigeria using limited data. *Data Brief.* (2020) 30:105683. doi: 10.1016/j.dib.2020.105683
51. Goldacre B, Harrison S, Mahtani K, Heneghan C. *WHO Consultation on Data and Results Sharing During Public Health Emergencies.* (2015). Available online at: https://www.who.int/medicines/ebola-treatment/background_briefing_on_data_results_sharing_during_phes.pdf?ua=1 (accessed June 18, 2020).
52. Aljunid SM, Srithamrongsawat S, Chen W, Bae SJ, Pwu R-F, Ikeda S, et al. Health-care data collecting, sharing, and using in Thailand, China Mainland, South Korea, Taiwan, Japan, and Malaysia. *Value in Health.* (2012) 15:S132–8. doi: 10.1016/j.jval.2011.11.004
53. Gewin V. Six tips for data sharing in the age of the coronavirus. *Nature.* (2020). doi: 10.1038/d41586-020-01516-0. [Epub ahead of print].


**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Copyright © 2021 Amit, Pepito, Gutierrez and Rawson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 13 September 2021
doi: 10.3389/fcimb.2021.594577

# Computer-Aided Medical Microbiology Monitoring Tool: A Strategy to Adapt to the SARS-CoV-2 Epidemic and That Highlights RT-PCR Consistency

*Linda Mueller*[1†], *Valentin Scherz*[1†], *Gilbert Greub*[1,2], *Katia Jaton*[1‡] *and Onya Opota*[1*‡]

[1] Insitute of Microbiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland,
[2] Infectious Diseases Service, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

OPEN ACCESS

**Edited by:**
*Max Maurin,*
*Université Grenoble Alpes, France*

**Reviewed by:**
*Meghan Starolis,*
*Quest Diagnostics, United States*
*Ahmad Qasem,*
*University of Central Florida, United States*

***Correspondence:**
*Onya Opota*
*Onya.Opota@chuv.ch*

[†]*These authors have contributed equally to this work*

[‡]*These authors jointly supervised the work*

**Specialty section:**
*This article was submitted to Clinical Microbiology, a section of the journal Frontiers in Cellular and Infection Microbiology*

**Received:** *13 August 2020*
**Accepted:** *17 August 2021*
**Published:** *13 September 2021*

**Citation:**
*Mueller L, Scherz V, Greub G, Jaton K and Opota O (2021) Computer-Aided Medical Microbiology Monitoring Tool: A Strategy to Adapt to the SARS-CoV-2 Epidemic and That Highlights RT-PCR Consistency. Front. Cell. Infect. Microbiol. 11:594577. doi: 10.3389/fcimb.2021.594577*

Since the beginning of the COVID-19 pandemic, important health and regulatory decisions relied on SARS-CoV-2 reverse transcription polymerase chain reaction (RT-PCR) results. Our diagnostic laboratory faced a rapid increase in the number of SARS-CoV-2 RT-PCR. To maintain a rapid turnaround time, we moved from a case-by-case validation of RT-PCR results to an automated validation and immediate results transmission to clinicians. A quality-monitoring tool based on a homemade algorithm coded in R was developed, to preserve high quality and to track aberrant results. We present the results of this quality-monitoring tool applied to 35,137 RT-PCR results. Patients tested several times led to 4,939 pairwise comparisons: 88% concordant and 12% discrepant. The algorithm automatically solved 428 out of 573 discrepancies. The most likely explanation for these 573 discrepancies was related for 44.9% of the situations to the clinical evolution of the disease, 27.9% to preanalytical factors, and 25.3% to stochasticity of the assay. Finally, 11 discrepant results could not be explained, including 8 for which clinical data was not available. For patients repeatedly tested on the same day, the second result confirmed a first negative or positive result in 99.2% or 88.9% of cases, respectively. The implemented quality-monitoring strategy allowed to: i) assist the investigation of discrepant results ii) focus the attention of medical microbiologists onto results requiring a specific expertise and iii) maintain an acceptable turnaround time. This work highlights the high RT-PCR consistency for the detection of SARS-CoV-2 and the necessity for automated processes to handle a huge number of microbiological results while preserving quality.

**Keywords: COVID-19, SARS-CoV-2, RT-PCR, biomedical validation, R-script, algorithm, quality-monitoring, delta checks**

## INTRODUCTION

The rapid spread of the COVID-19 pandemic caused unprecedented challenges for diagnostic microbiology laboratories. Rapid and high throughput SARS-CoV-2 reverse transcription polymerase chain reaction (RT-PCR) developed early during the crisis became the cornerstone of patient diagnosis as well as hospital and public health management (Caruana et al., 2020; Corman et al., 2020; Tadini et al., 2020). Consequently, microbiology laboratories were reorganized to respond to the high demand for SARS-CoV-2 testing (Posteraro et al., 2020). This situation required (i) the rapid adaptation of infrastructures, (ii) quick validation and implementation of new RT-PCR assays, (iii) working hour extension and new workforce employment. Yet, the quality of results provided by clinical microbiology laboratories, SARS-CoV-2 testing and routine analyzes, had to be maintained throughout the crisis.

Our molecular diagnostic laboratory located in a tertiary care university hospital faced a rapid increase in the number of SARS-CoV-2 PCR with up to 1,007 tests per days at the peak of the epidemic. Our analysis platform set was progressively extended from a high-throughput MDx platform, to the cobas 6800 system (introduced on 24.03.2020) and the Xpert Xpress SARS-CoV-2 assay (introduced on 21.04.2020) in response to the high volume of SARS-CoV-2 testing. Additionally, validation procedures had to be simplified in our laboratory. To ensure the best quality, two validation steps are usually applied prior to result transmission to clinicians: the technical validation of the assay followed by the biomedical validation of the results by medical microbiologists, who consider the specific clinical setting (Greub et al., 2015). Biomedical validation appeared as a bottleneck in the SARS-CoV-2 analytical workflow which could extend the turnaround time (TAT), with the risk of affecting clinical outcomes, infection prevention strategies and public health decisions (Hawkins, 2007). To maintain a minimal TAT, results were released to the clinicians after technical validation based on the FastFinder software (UgenTec NV, Hasselt, Belgium) that automatically analyzes RT-PCR amplification curves.

The limited experience in these newly implemented RT-PCR assays, including their performance (Kokkinakis et al., 2020), raised the need for an active surveillance of the quality of provided results. Delta checks are commonly used in clinical chemistry laboratories to monitor analytical throughputs that outreach capacity for sample-by-sample validation. "Delta checks" describes a process where discrepancies in sequential results of the same patient are detected to prompt repetition of the analysis (Schifman et al., 2017). We wondered whether a similar approach could be used to monitor the quality of SARS-CoV-2 results obtained in our laboratory. Thus, we developed a quality-monitoring methodology based on a homemade algorithm programmed in R to monitor SARS-CoV-2 RT-PCR results. Such methodology leveraged repeated testing to identify potential preanalytical or analytical culprits as well as cases requiring further biomedical investigations. The algorithm developed in-house aimed to restrict the list of discrepancies truly requiring investigation.

In this article, we present the results obtained through the application of our quality surveillance on data from the first four months of the COVID-19 crisis in our laboratory. Besides its role as a quality management tool, application of this surveillance allowed us to quickly gain knowledge about RT-PCR assays applied to a novel virus and new disease. In particular, this process allowed us to identify clinical specimen with significant added value (i.e. patients with unexpected discrepant results) and the presence of long-term carrying patients.

## MATERIALS AND METHODS

### RT-PCR and Samples

Samples collected from patients with suspected COVID-19 or for screening were tested by RT-PCR, using either our high-throughput MDx platform (Greub et al., 2015), the cobas SARS-CoV-2 qualitative test (Roche, Basel, Switzerland) and the Xpert SARS-CoV-2 test (Cepheid, California, USA). The E gene was targeted by the RT-PCR performed on the MDx platform (Greub et al., 2015), as described by Corman and colleagues (Corman et al., 2020). The cobas SARS-CoV-2 targeted the E gene as well as the ORF1a/b and was performed according to the manufacturer guidelines. Finally, the Xpert SARS-CoV-2 test targeted both the N and the E gene. The three methods displayed similar performances for the detection of SARS-CoV-2 from various clinical specimens and similar cycle threshold (Ct) value when positive (Lieberman et al., 2020; Moran et al., 2020; Opota et al., 2020; Poljak et al., 2020). Samples were mainly collected from the upper respiratory tract. However, other types of samples were also tested and are listed in **Supplementary Table S1**. Data collection and analysis

Data was collected during the first four months of the epidemic in Switzerland (12.02.2020-12.06.2020) and included all SARS-CoV-2 RT-PCR analyzes conducted at the Institute of Microbiology of the Lausanne University Hospital (CHUV). Samples were collected from symptomatic as well as asymptomatic patients. However, at the beginning of the pandemic only symptomatic patients were tested for SARS-CoV-2. From 25.04.2020, the screening strategy was extended to all patients admitted at our hospital, including the asymptomatics (Moraz et al., 2020). SARS-CoV-2 RT-PCR results and basic contextual information were extracted from our Laboratory Information System (LIS) (MOLIS, CGM) and analyzed with R (Team, RC 2019) (version 3.6.1) language helped by packages from the *Tidyverse* (Wickham et al., 2019) environment.

### Discrepant Cases Identification and Classification

A R script was developed to automatically identify and classify discrepant cases. In this script, all analyzes from patients with multiple samples were compared to their previous results in a pairwise approach. Sample comparisons were then categorized as concordant or discrepant. Only discrepant results were further

processed. These discrepancies (positive versus negative or conversely) between consecutive samples were classified based on i) Ct values, ii) samples types and iii) reception dates (**Supplementary Figure S1**). Based on these records, discrepancies were classified by the algorithm as described in **Supplementary Table S2**. The script was designed to compare each sample only to the last relevant result. As described in **Supplementary Table S2** (Patient A), a positive nasopharyngeal swab followed by a negative PCR in blood, and then later by another positive nasopharyngeal swab, will lead to only one discrepancy: the negative blood classified as a "Low yield" sample. The second nasopharyngeal swab will be classified as concordant with the previous nasopharyngeal swab. Of note, two nasopharyngeal samples taken more than 10 days apart once negative and once positive with a Ct > 35, would be classified as "Stochastic" (Patients B and D) and not as "Time delay" (as it is for Patient C). Indeed, when none of these criteria are met (Sample type is not a "Low Yield", Ct are <35 and Delta Time between samples is <10 days) the result is classified as "To be investigated" (Patient E). Of note, discrepancies were classified according to the first matching criteria in the following order: "Low yield", "Stochastic", "Time delay" and "To be investigated". Furthermore, a result from a patient with three samples or more can be involved in a concordant and a discrepant pairwise comparisons. Indeed, the second of his analyzes could be in agreement with the first result but discrepant with the third. This decisional algorithm is graphically represented in **Supplementary Figure S1**. Code of this algorithm is available on https://github.com/valscherz/SARS-CoV-2_discrepant_screen.

Discrepant samples classified as "To be investigated" by the algorithm were then manually investigated, classified and assigned a putative explanation for the observed discrepancy (**Supplementary Table S3**). In this manual analysis, a discrepancy between two nasopharyngeal swabs taken within the same period (< 24h) and collected in different units or different hospitals (compatible with differences in sampling quality) were imputed to "Sample quality". When sampling sources were different (i.e. comparing an upper respiratory tract sample with a rectal swab), discrepancies were imputed to "Different sample types". Discrepancies between samples collected less than 10 days apart but with indications in clinical records supporting a recent infection or recent recovery were classified as "Clinical context". Finally, discrepancies compatible with none of these putative explanations were classified as "Unsolved". For visualization purposes, classified discrepancies were grouped into corresponding testing phases or context: clinical context, preanalytical or stochastic (**Supplementary Tables S3** and **S4**).

RT-PCR analyses were not repeated on the discrepant samples.

# RESULTS

## Post-Analytic Surveillance of SARS-CoV-2 RT-PCR Results

Since the implementation of SARS-CoV-2 RT-PCR assays and for a period of four months, 30,198 patients were tested by RT-PCR at the Institute of Microbiology of the Lausanne University Hospital (CHUV). This corresponded to 35,137 samples, among which 4,545 (12.9%) returned a positive result, whereas 30,592 (87.1%) were negative. Upper respiratory tract (URT) samples represented 98% of the tested specimens (**Supplementary Table S1**). The peak of number of analyzes took place on March 18th with up to 1,007 analyzes processed during the same day (**Figure 1**). The developed algorithm allowed the laboratory to process 3,214 patients having at least two specimens, as detailed in the **Figure 2**.

## Algorithm-Based and Manual Biomedical Investigation of Discrepancies

Our pipeline significantly reduced the number of discrepancies requiring human investigation and a probable explanation could be identified for most of the discrepant results. Indeed, 75% (n=428 of total 573) of the discrepant pairwise comparisons could be automatically attributed by the pipeline to a putative explanation, i.e., "stochastic", "low yield" or "time delay" (**Figures 3A**, **B** and **Supplementary Table S3**). Only the remaining discrepancies (n=145) did not fit any of the solving rules encoded in the algorithm and required investigations based on the available analytical and clinical information (**Supplementary Table S4**).

The profiles of putative explanations for discrepancies evolved depending on the time interval between the compared analyzes (**Figure 3C**). In samples received during the same day, our assessment explained 77.3% (n=34/44) of the discrepancies as related to the preanalytical phase (i.e. explained by the sample type or collection in different health centers) and 22.7% (n=10/44) to stochasticity of the RT-PCR reaction (Ct value >35). The discrepancies in results for samples received 1-3 days apart were explained by factors affecting the preanalytical phase (55.4%, n=51/92), followed by stochasticity (32.6%, n=30/92).

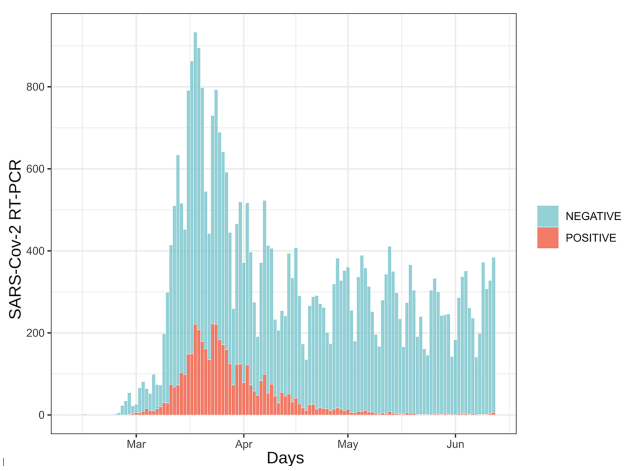



**FIGURE 1** | Daily number of SARS-CoV-2 RT-PCR assays. 35,137 analyzes are represented here. Samples are distributed according to their reception date. Blue bars represent samples for which RT-PCR results were negative (88%) while red bars depict positive samples (12%).

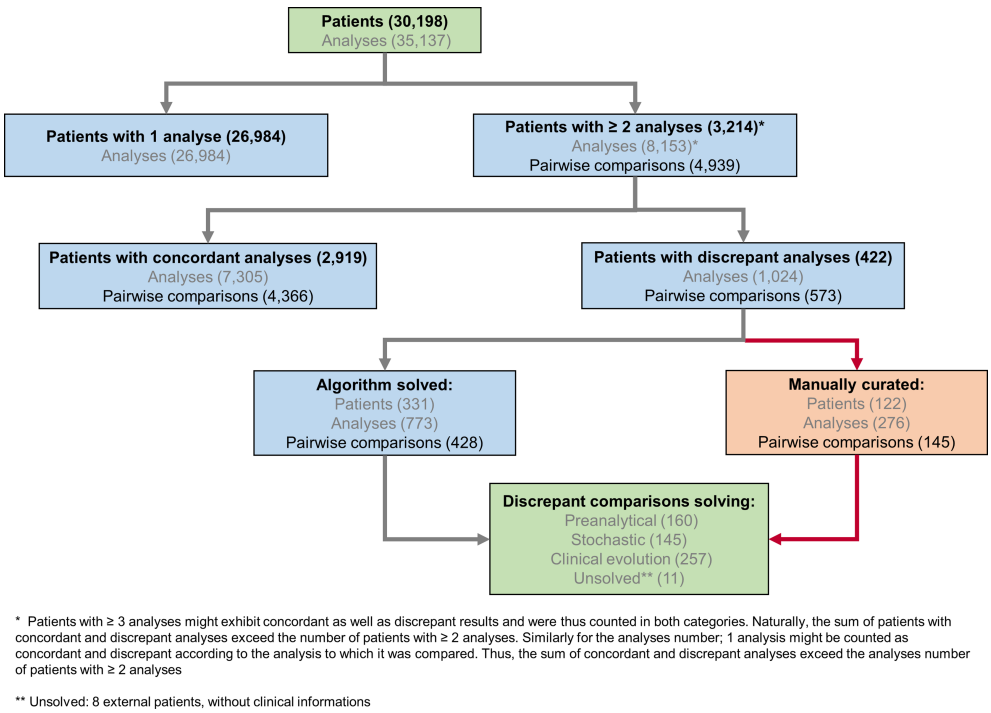



**FIGURE 2** | SARS-CoV-2 analytical flowchart. An R-based script was used to identify 3,214 patients with multiple, potentially discrepant, results (upper part of the chart) and therefore ensure their surveillance. Among these patients with multiple analyzes, 2,792 exhibited only concordant results, while 422 presented at least one discrepant pair of results. Of note, 127 patients had concordant as well as discrepant results; these latter were further processed by the algorithm as patients with discrepant results. When considering 4,939 pairwise comparisons of successive results for patients with multiple tests, 4,366 (88%) were concordant and 573 (12%) were discrepant. The algorithm processed discrepant results further and a potential explanation for the observed discrepancy was attributed either automatically or manually (lower part of the chart and **Supplementary Figure S1**).

Interestingly, 7 of the discrepancies observed in the 1-3 days interval were explainable by nosocomial (n=6) or community (n=1) acquired infections based on health records, which explained the quick negative to positive transition. These 7 discrepancies were thus classified in the clinical evolution context. As for the 4 remaining discrepancies, clinical records were not available for 3 and the last one remained unexplained. Investigation of discrepancies between samples received 4-10 days apart again incriminated mainly the preanalytical phase (41.6%, n=55/132), followed by stochasticity (30.3%, n=40/132). As expected, the discrepancies imputable to the clinical evolution of the disease based on clinical records (new infection or infection resolution) was greater in the 4-10 days interval since it represented 22.7% of the discrepancies (n=30/132). The 7 remaining discrepancies in this time interval could not be explained, either in absence (n=5) or in presence (n=2) of clinical information. Over 10 day, the clinical evolution of the disease was the main explanation (72.1%, n=220/305) for discrepancies, as it was the default explanation retained by our automatic pipeline for discrepant results from samples collected more than 10 days apart in absence of any other explanation.

In the overall assessment of the 573 discrepancies from samples taken up to 90 days apart, 44.9% (n=257) of discrepancies could be explained by the clinical evolution of the disease (e.g. indications in clinical records for new contagion, time delay making new infection or infection resolution likely) (**Figure 3D**). 27.2% (n=160) of cases had arguments for factors incriminating the preanalytical phase (discrepant results among samples collected by different health centers, inclusion of samples rarely positive as blood); and 25.3% (n=145) of the discrepant comparisons could be explained by analytical stochasticity in presence of low RNA loads (Ct value > 35 for the positive sample followed or preceded by a negative sample). No clear explanation could be identified for 1.9% (n=11) of the discrepancies (classified as "Unsolved"). Among the unsolved situations, 8 samples were submitted to our laboratory by external care centers or private laboratories and clinical records were thus not accessible. No explanation for discrepancies could be found for 3 cases, despite the availability of full clinical documentation. Moreover, short-term negative to positive transitions were compatible with 21 nosocomial and 8 community-acquired infections based on clinical records (**Supplementary Table S4**).

## Evolution of Discrepancy Patterns Across the Epidemic Period

The pattern in transitions (negative result followed by a positive result or the reverse) among discrepancies evolved over the

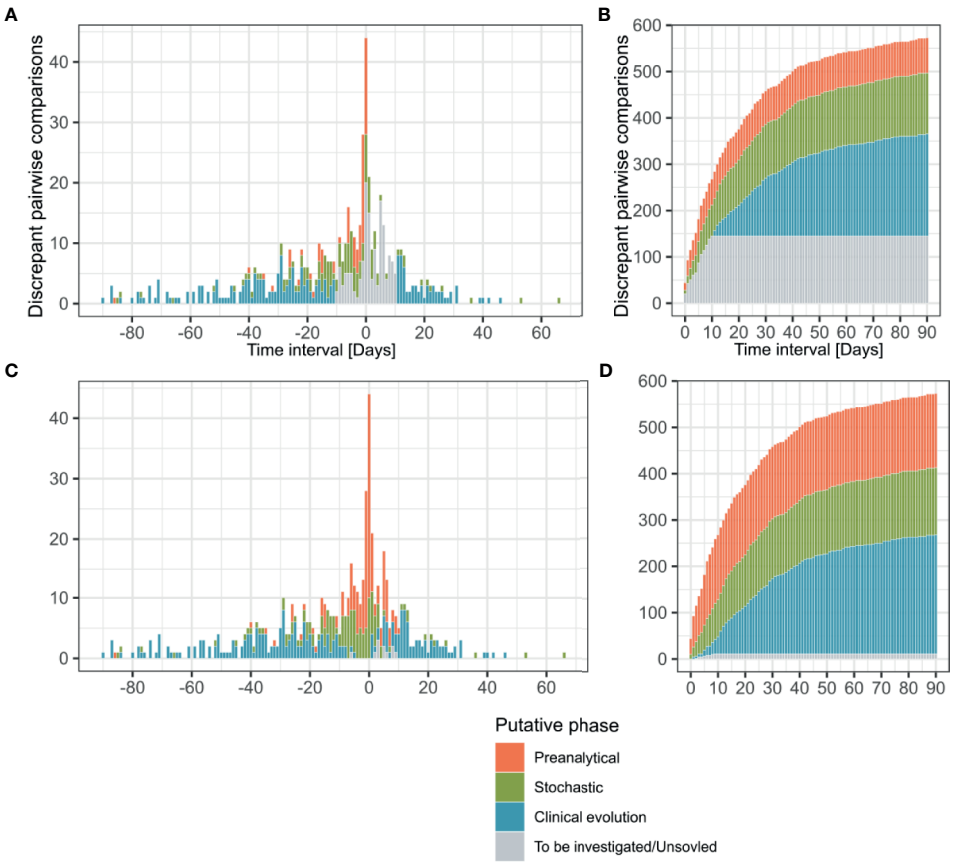



**FIGURE 3** | Time interval between discrepant analyzes, putative phase assignment and cumulative curves. Pairwise comparisons (n = 573) are represented according to the interval between their reception dates and colored depending on the analytical phase which best explained the observed discrepancy. Before manual curation discrepancies were classified by the algorithm as "Preanalytical", "Stochastic", "Clinical evolution" and "To be investigated" **(A, B)**. After manual curation comparisons previously categorized as "To be investigated" were reassigned to the same categories or as "Unsolved" **(C, D)**. Transitions from a negative to a positive result are represented in the positive side of x axe, while positive to negative transitions are plotted on the negative side **(A, C)**.

studied period of four months that correspond to the four first months of epidemic in our region. In the first two months (12.02-12.04.2020), 71.3% of observed transitions were negative to positive (n=154/216). Conversely, in the last two months (13.04-12.06.2020), 81.2% of the transitions went from a positive to a negative result (n=290/357), which contributed to an overall trend of 61.4% of positive to negative discrepancies (n = 352/573) (**Figures 3A, C**). Such observation was expected and imputable to follow-up of patients with resolving infections.

## Sustained RT-PCR Positive Results in Patients

Besides discrepancies, we also investigated sustained positivity in patients. In our analysis, the longest time interval between two positive results from the same patient was of 83 days (**Figures 4A, B** and **5**). Considering the time interval between their first and the last positive result, we observed 32, 11, and 3 patients with sustained positivity in samples taken over 30, 50 and 70 days apart, respectively. This observation questions the presence of active viral replication or only of viral traces remaining from the resolving infection. In line, the last sample of these long-time carriers displayed a low viral load with Ct values around 35 (corresponding to 3,800 copies/ml) in all but one notable exception (**Figure 5**).

## Pairwise Analyzes Highlight The Consistency of SARS-CoV-2 RT-PCR Results

Reports questioned the performance of RT-PCR for SARS-CoV-2 detection, which supported recommendations for repeated testing (13). In our dataset, only 2.5% of the negative results obtained from an URT sample (n=733/29,714) were followed by an additional analysis 1 hour to 3 days after initial testing. In comparison, 0.6% (n=28/4,451) of positive results from URT samples were followed by a second analysis over the same time interval. Thus, if repeated testing remained limited, a negative result was still significantly more often challenged than a positive result by clinicians (Pearson's Chi-squared Test, $p < 0.001$, OR = 4.0).

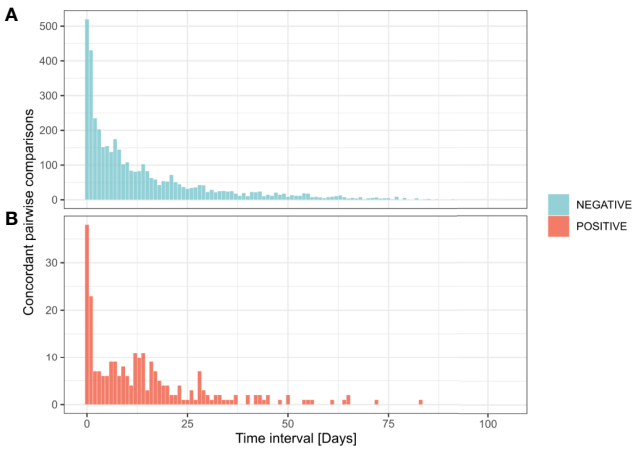



**FIGURE 4** | Time interval between successive concordant analyzes from the same patient. Distribution of 4,366 pairwise concordant comparisons according to time interval between their reception dates. Negative concordant comparisons for SARS-CoV-2 detection (n = 4,116) **(A)**. Positive concordant comparisons for SARS-CoV-2 detection (n = 250) **(B)**.

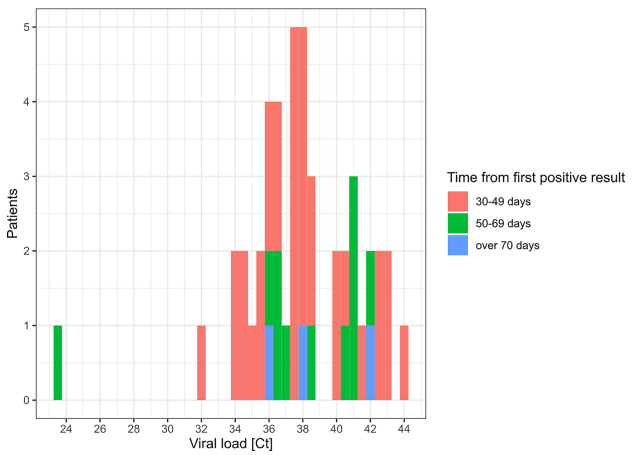



**FIGURE 5** | Distribution of Ct values for patients with sustained RT-PCR positive results. This histogram represents the Ct value (maximal if more than 1) for SARS-CoV-2 detection by RT-PCR in the last positive sample of patients with sustained positivity in samples taken 30 days apart or more.

An evaluation of the performance of the RT-PCR for SARS-CoV-2 detection based on samples collected over a short time interval showed a good level of concordance. Indeed, an initial negative result in an URT sample was confirmed in 99.2% (n=243/245) of cases for patients tested twice on the same day (**Table 1**); both discrepancies could be explained by stochasticity since associated to high Ct values. Conversely, a first positive result was confirmed in 88.9% (n=24/27) of cases; two discrepancies could be explained by stochasticity too, while the third involved a positive nasopharyngeal swab and a negative throat swab, a sample site shown to be less sensitive for SARS-CoV-2 detection (14). As

**TABLE 1** | Agreement between URT samples collected in a repeatedly in patients over a short time-interval.

| | Same day (n = 272) | 1-3 day (n = 826) | 4-6 days (n = 489) | 7-10 days (n = 585) |
|---|---|---|---|---|
| Concordant Negative results | 243 | 750 | 430 | 516 |
| Concordant Positive results | 24 | 21 | 7 | 20 |
| Negative to Positive transition | 2 | 46 | 38 | 28 |
| Positive to Negative transition | 3 | 9 | 14 | 21 |
| **Concordant negative agreement** | 99.2% | 94.2% | 91.9% | 94.9% |
| **Concordant positive agreement** | 88.9% | 70% | 33.3% | 48.8% |

expected, concordance rates diminished over time: negative result concordance for URT samples went from 99.2% for samples collected along the same day to 94.2% for samples collected 1-3 days apart. On the same time intervals, the concordance for positive results evolved from 88.9% to 70.0%.

## DISCUSSION

This work presents the importance of a homemade algorithm developed in response to the need for quality surveillance of SARS-CoV-2 RT-PCRs which throughput exceeded the ability to conduct manual biomedical validation for each sample. Applied to the 35,137 SARS-CoV-2 RT-PCRs performed in our diagnostic laboratory from February 12 to June 12, 2020, the algorithm identified 3,214 patients owing multiple tests. These patients represented an opportunity for quality assessment of our analyzes, but also required careful attention to investigate potential discrepancies. Among the 3,124 patients tested multiple times, we observed a majority (86.8%) of concordant results, mostly negative (96.8%). Of these patients, 422 exhibited at least one pair of discrepant results. Together, the clinical evolution of the disease (44.9%), preanalytical factors (27.9%) and stochasticity around the limit of detection (25.3%) were the most likely explanations retained for the 573 observed discrepancies. Only 1.4% of the cases remained unexplained because clinical records were not available. Despite availability of all records, 0.5% of the results remained unexplained.

Expectedly, the natural evolution of the disease explained most of the observed discrepancies. The preanalytical factors were the second most frequent source of discrepancies. This observation is in line with previous reports that described this testing phase as an important source of errors in general in clinical laboratories (Plebani, 2006; West et al., 2017). Stochasticity was the only identified source of discrepancy directly related to the analytical phase. However, the clinical impact of these discrepancies could be limited since low viral loads are expected in the late course of the disease, at time when the infectivity might be diminished (Jacot et al., 2020; Moraz et al., 2020; Yu et al., 2020). Of note, samples were analyzed according to the laboratory workflow. Indeed, samples belonging to the same patient were not systematically analyzed with the same method. Thus, the difference in LOD proper to each assay

might also be a source of the discrepancies classified as stochastic (Opota et al., 2020).

Sustained positivity with high Ct values was observed in 45 patients (Fig. 5). According to studies focusing on prolonged presence of viral nucleid acid, viral traces might not be associated with effective infectiousness and mostly correspond to nonculturable samples (Hong et al., 2020; Huang et al., 2020; Wolfel et al., 2020). Nevertheless, many factors can impact the viral load in a clinical specimen (i.e. quality of the sampling); therefore, several co-variables have to be taken into consideration to address the contagiousness (Jacot et al., 2020; Moraz et al., 2020).

Our results support the good performance of RT-PCR in URT samples for SARS-CoV-2 detection. Nevertheless, these agreement rates should be considered with caution, particularly due to the small number of positive samples for which additional testing was requested. Another limitation of our work is that while we intended to use an unbiased algorithm stable over time to investigate discrepancies in results, some of the applied criteria were partly arbitrary (e.g. the 10 days limit to consider discrepancies as due to the clinical evolution of the disease). Furthermore, our process retained a single explanation for each observed discrepancy, while more could be applicable. While arguable, these choices were made to fit a strategy of quality monitoring. Indeed, the primary aim of the present methodology was to attribute the observed discrepancies to the most likely explanation to focus on truly unexplainable and clinically problematic cases. Clinical laboratory vulnerabilities during the COVID-19 pandemic were the subject of a recent publication by Lippi et al. (Lippi and Plebani, 2020). Our assessment overlaps with some of the preanalytical culprits identified by the authors such as specimen collection (see "detailed explanations" **Supplementary Tables S3** and **S4**). However, some other potential vulnerabilities were not considered as probable causes for discrepancies in our assessment, since they are covered by other pre-existing quality management procedures in our laboratory. For instance, samples missing patient identification were systematically rejected. Moreover, internal extraction controls and amplification controls were systematically included to detect samples that might contain interfering substances compromising the amplification (Poljak et al., 2020).

This is the first implementation in the clinical microbiology facility of the Lausanne University Hospital of a quality monitoring tool resembling a "Delta check" applied in clinical chemistry laboratories (Schifman et al., 2017). "Delta checks" are usually restricted to analytes exhibiting limited short-term variations and is as such unsuitable to microbiology results. Yet, helped by an algorithm capable of considering expected variations in results, we could adapt the concept of "Delta check" to SARS-CoV-2 RT-PCR. Similar longitudinal observation of results and algorithm-based selection of "cases to investigate" could also be applied to other high-throughput microbiology laboratory assays, either by the implementation of an ad-hoc software as presented here or by rules embedded in the LIS.

The strategy applied for the management of large amount of SARS-CoV-2 samples in our center, comprising the extension of our analysis platform set and the introduction of an automatic validation, allowed to reduce the median TAT from 6.9h to 4.8h between February 24$^{th}$ and June 9$^{th}$, 2020 (Marquis et al., 2021).

However, the duration of the biomedical validation step, depending on the pathogen and the epidemiological situation, remains to be assessed.

This work emphasized the benefit of an automated algorithm capable of finding discrepant results and attributing them to corresponding testing phases. This computer-aided methodology outlines that besides the expected evolution of the disease, most of discrepant results are compatible with preanalytical factors. Moreover, most of URT samples collected repeatedly in a short timeframe showed consistent results, displaying the good reproducibility of the RT-PCR for SARS-CoV-2 detection. Application of this method for quality monitoring enabled to focus on problematic cases requiring biomedical expertise while maintaining an acceptable TAT.

## DATA AVAILABILITY STATEMENT

The code of the algorithm presented in this article can be found under https://github.com/valscherz/SARS-CoV-2_discrepant_screen.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

VS and LM conducted the analyses under GG, KJ, and OO supervision for study design and data interpretation. VS designed the R script. LM investigated discrepant cases. LM and VS wrote the original manuscript. All authors contributed to the article and approved the submitted version.

## ACKNOWLEDGMENTS

We would like to deeply thank all the staff of the Institute of Microbiology of the Lausanne University Hospital. In particular, we thank all the staff of the Laboratory of Molecular Diagnostic of the Institute of Microbiology of the University of Lausanne and all the training. This article was submitted as a preprint to MedRxiv as Mueller et al. 2020.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2021.594577/full#supplementary-material

## REFERENCES


Caruana, G., Croxatto, A., Coste, A. T., Opota, O., Lamoth, F., Jaton, K., et al. (2020). Diagnostic Strategies for SARS-CoV-2 Infection and Interpretation of Microbiological Results. *Clin. Microbiol. Infect.* 26 (9), 1178–1182. doi: 10.1016/j.cmi.2020.06.019

Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., et al. (2020). Detection of 2019 Novel Coronaviru-Ncov) by Real-Time RT-PCR. *Euro Surveill* 25, 25–30. doi: 10.2807/1560-7917.ES.2020.25.3.2000045

Greub, G., Sahli, R., Brouillet, R., and Jaton, K. (2015). Ten Years of R&D Anf Full Automation in Molecular Diagnosis. *Future Microbiol.* 11, 403–425. doi: 10.2217/fmb.15.152

Hawkins, R. (2007). Laboratory Turnaround Time. *Clin. Biochem. Rev.* 28, 179–194.

Hong, K., Cao, W., Liu, Z., Lin, L., Zhou, X., Zeng, Y., et al. (2020). Prolonged Presence of Viral Nucleic Acid in Clinically Recovered COVID-19 Patients was Not Associated With Effective Infectiousness. *Emerg. Microbes Infect.* 9, 2315–2321. doi: 10.1080/22221751.2020.1827983

Huang, C. G., Lee, K. M., Hsiao, M. J., Yang, S. L., Huang, P. N., Gong, Y. N., et al. (2020). Culture-Based Virus Isolation To Evaluate Potential Infectivity of Clinical Specimens Tested for COVID-19. *J. Clin. Microbiol.* 58. doi: 10.1128/JCM.01068-20

Jacot, D., Greub, G., Jaton, K., and Opota, O. (2020). Viral Load of SARS-CoV-2 Across Patients and Compared to Other Respiratory Viruses. *Microbes Infect.* 22, 617–621. doi: 10.1016/j.micinf.2020.08.004

Kokkinakis, I., Selvy, K., Favrat, B., Genton, B., and Cornuz, J. (2020). Performance Du Frottis Nasopharyngé-PCR Pour Le Diagnostic Du Covid-19 Recommandations Pratiques Sur La Base Des Premières Données Scientifiques. *Rev. Med. Suisse* 16, 699–701.

Lieberman, J. A., Pepper, G., Naccache, S. N., Huang, M. L., Jerome, K. R., and Greninger, A. L. (2020). Comparison of Commercially Available and Laboratory-Developed Assays for In Vitro Detection of SARS-CoV-2 in Clinical Laboratories. *J. Clin. Microbiol.* 58, e00821–20. doi: 10.1128/JCM.00821-20

Lippi, G., and Plebani, M. (2020). The Critical Role of Laboratory Medicine During Coronavirus Disease 2019 (COVID-19) and Other Viral Outbreaks. *Clin. Chem. Lab. Med.* 58, 1063–1069. doi: 10.1515/cclm-2020-0240

Marquis, B., Opota, O., Jaton, K., and Greub, G. (2021). Impact of Different SARS-CoV-2 Assays on Laboratory Turnaround Time. *J. Med. Microbiol.* 70, 1–6. doi: 10.1099/jmm.0.001280

Moran, A., Beavis, K. G., Matushek, S. M., Ciaglia, C., Francois, N., Tesic, V., et al. (2020). Detection of SARS-CoV-2 by Use of the Cepheid Xpert Xpress SARS-CoV-2 and Roche Cobas SARS-CoV-2 Assays. *JCM* 58, e00772–e00720. doi: 10.1128/JCM.00772-20

Moraz, M., Jacot, D., Papadimitriou-Olivgeris, M., Senn, L., Greub, G., Jaton, K., et al. (2020). Universal Admission Screening Strategy for COVID-19 Highlighted the Clinical Importance of Reporting SARS-CoV-2 Viral Loads. *New Microbes New Infect.* 38, 100820. doi: 10.1016/j.nmni.2020.100820

Mueller, L., Scherz, V., Greub, G., Jaton, K., and Opota, O. (2020). Computer-Aided Medical Microbiology Monitoring Tool: A Strategy to Adapt to the SARS-CoV-2 Epidemic and That Highlights RT-PCR Consistency. *MedRxiv* [Preprint]. doi: 10.1101/2020.07.27.20162123

Opota, O., Brouillet, R., Greub, G., and Jaton, K. (2020). Comparison of SARS-CoV-2 RT-PCR on a High-Throughput Molecular Diagnostic Platform and the Cobas SARS-CoV-2 Test for the Diagnostic of COVID-19 on Various Clinical Samples. *Pathog. Dis.* 78 (8). doi: 10.1093/femspd/ftaa061

Plebani, M. (2006). Errors in Clinical Laboratories or Errors in Laboratory Medicine? *Clin. Chem. Lab. Med.* 44, 750–759. doi: 10.1515/CCLM.2006.123

Poljak, M., Korva, M., Knap Gasper, N., Fujs Komlos, K., Sagadin, M., Ursic, T., et al. (2020). Clinical Evaluation of the Cobas SARS-CoV-2 Test and a Diagnostic Platform Switch During 48 Hours in the Midst of the COVID-19 Pandemic. *J. Clin. Microbiol.* 58 (6), e00599–20. doi: 10.1128/JCM.00599-20

Posteraro, B., Marchetti, S., Romano, L., Santangelo, R., Morandotti, G. A., Sanguinetti, M., et al. (2020). Clinical Microbiology Laboratory Adaptation to COVID-19 Emergency: Experience at a Large Teaching Hospital in Rome, Italy. *Clin. Microbiol. Infect.* 26 (8), 1109–1111. doi: 10.1016/j.cmi.2020.04.016

Schifman, R. B., Talbert, M., and Souers, R. J. (2017). Delta Check Practices and Outcomes: A Q-Probes Study Involving 49 Health Care Facilities and 6541 Delta Check Alerts. *Arch. Pathol. Lab. Med.* 141, 813–823. doi: 10.5858/arpa.2016-0161-CP

Tadini, E. P.-O., Opota, M., Moulin, O., Lamoth, E., Manuel, F., Lhopitallier, O., et al. (2020). SARS-CoV-2, Un Point Dans La Tourmente. *Rev. Med. Suisse* 16 (692), 917–923.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (Vienna, Austria). Available at: https://www.R-project.org/.

West, J., Atherton, J., Costelloe, S. J., Pourmahram, G., Stretton, A., and Cornes, M. (2017). Preanalytical Errors in Medical Laboratories: A Review of the Available Methodologies of Data Collection and Analysis. *Ann. Clin. Biochem.* 54, 14–19. doi: 10.1177/0004563216669384

Wickham, H., Averick, M., Bryan, J., Chang, W., Mcgowan, L., François, R., et al. (2019). Welcome to the Tidyverse. *J. Open Source Software* 4 (43). doi: 10.21105/joss.01686

Wolfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Muller, M. A., et al. (2020). Virological Assessment of Hospitalized Patients With COVID-2019. *Nature* 581, 465–469. doi: 10.1038/s41586-020-2196-x

Yu, F., Yan, L., Wang, N., Yang, S., Wang, L., Tang, Y., et al. (2020). Quantitative Detection and Viral Load Analysis of SARS-CoV-2 in Infected Patients. *Clin. Infect. Dis.* 71 (15), 793–798. doi: 10.1093/cid/ciaa345


**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2021 Mueller, Scherz, Greub, Jaton and Opota. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*

# Advantages of publishing in Frontiers

**OPEN ACCESS**
Articles are free to read for greatest visibility and readership

**FAST PUBLICATION**
Around 90 days from submission to decision

**HIGH QUALITY PEER-REVIEW**
Rigorous, collaborative, and constructive peer-review

**TRANSPARENT PEER-REVIEW**
Editors and reviewers acknowledged by name on published articles

**Frontiers**
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

**Visit us:** www.frontiersin.org
**Contact us:** frontiersin.org/about/contact

**REPRODUCIBILITY OF RESEARCH**
Support open data and methods to enhance research reproducibility

**DIGITAL PUBLISHING**
Articles designed for optimal readership across devices

**FOLLOW US**
@frontiersin

**IMPACT METRICS**
Advanced article metrics track visibility across digital media

**EXTENSIVE PROMOTION**
Marketing and promotion of impactful research

**LOOP RESEARCH NETWORK**
Our network increases your article's readership