Optimizing Healthcare Delivery in Resource-Limited Settings

by

Emma Gibson

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author . . . .
Sloan School of Management
August 1, 2022

Certified by. . . . .
Jónas Oddur Jónasson
Associate Professor in Operations Management
Thesis Supervisor

Accepted by . . . .
Georgia Perakis
William F. Pounds Professor of Management Science
Co-director, Operations Research Center

Optimizing Healthcare Delivery in Resource-Limited Settings

by

Emma Gibson

Submitted to the Sloan School of Management
on August 1, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In resource-limited settings, critical diagnostic testing services are frequently provided through hierarchical networks comprised of healthcare facilities that collect diagnostic samples (e.g. blood, nasal swabs) from patients, and centralized medical laboratories that analyze these samples. The first part of this thesis focuses on diagnostic sample transportation systems, which are used to move samples and test results between various locations within centralized networks.

In Chapter 2, we describe the design and implementation of a low-cost information sharing system which allows healthcare workers to report daily sample volumes at each facility within the network using a simple text-based interface accessible on any standard mobile phone. The feasibility and effectiveness of this system were assessed in a field trial at 51 healthcare facilities in Malawi, which achieved high rates of participation and accuracy.

In Chapter 3 we propose an *optimized sample transportation* system which uses data reported by healthcare facilities to generate efficient routes for sample couriers on a daily basis. This system was implemented in three districts in Malawi, where it reduced average transportation delays by 25% and decreased the proportion of unnecessary trips by 55%.

In Chapter 4 we evaluate operational strategies for the deployment of *Point-of-Care* (POC) testing at healthcare facilities in Malawi. We develop a mixed-integer model to optimize the allocation of POC instruments to strategic locations within the diagnostic network in order to maximize the benefits of viral load monitoring services for people living with HIV. Our analysis indicates that the most effective POC deployment policies include a combination of *targeted* POC testing of high-risk patients, as well as capacity-sharing strategies such as near-POC testing.

In Chapter 5, we study *survival analysis* models, which are frequently used to analyze health outcomes and identify risk factors associated with morbidity and mortality. We present a new Globally Optimized Survival Trees algorithm that leverages mixed-integer optimization and local search techniques to generate interpretable survival tree models. We demonstrate that this algorithm improves on the accuracy of existing survival tree methods, particularly in large datasets.

Thesis Supervisor: Jónas Oddur Jónasson
Title: Associate Professor in Operations Management

Acknowledgments

During my time at MIT, I have had the opportunity to work on challenging and exciting research in an area that is very meaningful to me. I would like to thank my thesis advisor, Jónas Jónasson, for providing these opportunities and encouraging me to pursue them over the past four years. Jónas has been an incredibly supportive advisor, and I am especially grateful for his trust and flexibility during our work together, which helped me to become a more confident and creative researcher. I admire both his commitment to his work and the thoughtful way that he approaches every meeting and collaboration, and it has been a great privilege to work with Jónas during my PhD.

I would like to thank Georgia Perakis and Joann de Zegher for serving on my PhD thesis committee. Joann's research has been a source of inspiration during my PhD, and I was fortunate to benefit from her experience and insight on my committee. Georgia has been a fantastic mentor since my first year in the ORC and has helped me to navigate challenging periods during my PhD. I am grateful for her advice and support, as well as all of the time and effort that she has contributed to the ORC.

I am grateful to Dimitris Bertsimas, who played a significant role in my decision to join the ORC and supported me during my first PhD research project and my qualifying exams. I would like to thank Colin Fogarty and Rob Freund for giving me the opportunity to TA for the *Analytics Edge* and for their flexibility and support during the pandemic-related disruptions. I am also grateful to Colin for chairing my General Exam committee.

I would like to thank many external collaborators who have played a significant role in my research over the last few years. Sarang Deo has brought a great deal of experience and insight to our work on diagnostic networks, and I am grateful for his patience, advice and feedback. Kara Palamountain played a key role in initiating many productive research collaborations and was a great source of support in the sample transportation work. I would also like to thank Andrew Phillips for sharing the HIV Synthesis Model and providing valuable advice on the Point-of-Care testing

project.

Mphatso Kachule has been a fantastic collaborator in our work on sample transportation networks. It was a privilege to work with Riders 4 Health Malawi and I am grateful to all of the staff who have contributed to the implementation and ongoing development of the ST optimization system, particularly Denview Magalasi, Emmanuel Ngwira, Innocent Guta, and Zwelithini Golowa.

I would like to thank past and present ORC students who have been great friends and collaborators over the last few years. I am grateful to Agni Orfanoudaki for her support at stressful times, as well as for the many happy memories of fun times. I am grateful to Daniel Killian for his thoughtfulness and encouragement during our work together, and to Phil Chodrow and Jazmin Furtado for many years of friendship.

It has been an honor to work with the ORC REFS, including Sebastien Martin, Arthur Delarue, Julia Yan, Kayla Cummings, Lily Liu, and Sam Gilmour. Each of the REFS has made my time at MIT better through their kindness and encouragement, and I am grateful for all the time and energy that they contribute to supporting other students in the ORC.

Finally, I would like to thank my family. I am grateful to my brothers, James and Matthew; my parents-in-law, Chery and Duncan, and especially my parents, Elsbeth and Roger. I appreciate their incredible love and support, and all that they have done to help me to reach this milestone. Most of all, I would like to thank Alastair for joining me on this journey as my friend, my family, and so much more.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Healthcare systems in resource-limited settings face substantial operational challenges including insufficient funding, poor infrastructure, skills shortages, and unreliable supply chains. In sub-Saharan Africa, these challenges are compounded by high rates of communicable disease which increase the demand on under-resourced systems. The global HIV epidemic is a stark example of the disproportionate impact of disease in populations with limited access to healthcare—two-thirds of people living with HIV are located in sub-Saharan Africa, and HIV prevalence among adults is as high as 20% in some countries in the region.

The first part of this thesis focuses on *sample transportation* (ST) systems, which play a crucial role in providing access to centralized diagnostic testing services for people living with HIV. ST systems are used to move medical samples (e.g., blood, sputum) between healthcare facilities and laboratories, and to facilitate the delivery of test results to healthcare facilities. Chapter 2 describes the design and implementation of a low-cost data sharing platform to monitor sample volumes at healthcare facilities in Malawi, while Chapter 3 presents an optimization algorithm for scheduling and routing sample transport couriers in Malawi's national diagnostic network.

In Chapter 4 we evaluate the potential impact of introducing *Point-of-Care* (POC) technology at healthcare facilities to provide faster and more reliable diagnostic testing. We develop detailed cost and impact models to optimize the deployment of POC instruments within the existing diagnostic network, and assess various operational strategies to improve the cost-effectiveness of POC testing.

Chapter 5 focuses on *survival analysis* models, which are frequently used to analyze health outcomes in observational studies and to understand risk factors associated with morbidity and mortality. We present a new algorithm to generate optimized, interpretable decision-tree models for survival data.

1.1 Improving Data Visibility in Diagnostic Networks

Centralized diagnostic networks play a key role in healthcare systems in resource-limited settings, providing low-cost, high-volume diagnostic testing to facilitate the treatment of HIV, Tuberculosis (TB), and other diseases. In these hierarchical networks, medical samples are collected from patients at healthcare facilities—often located in remote or rural areas—and transported to centralized laboratories for testing. These systems frequently operate settings which lack the necessary communications infrastructure to share data and information between different levels of the network, resulting in significant operational inefficiencies in the transportation of diagnostic samples and results.

Chapter 2 describes a distributed data-collection system which leverages basic mobile phone technology to gather reports on the quantity and location of diagnostic samples requiring transportation within Malawi's diagnostic network. The system is based on an Unstructured Supplementary Service Data (USSD) application that enables health workers at remote facilities to submit daily sample volume reports from their personal mobile phones.

We assess the feasibility, adoption, and accuracy of this system through a field trial conducted at 51 healthcare facilities in Malawi. During the study period, healthcare facility staff submitted 37,771 reports and sustained average daily participation rates of approximately 80%. The average accuracy of submitted reports ranged from 81% to 89% for different disease programs. These results demonstrate that the USSD system is an effective, low-cost tool to facilitate information sharing within geographically dispersed diagnostic networks.

Joint work with S. Deo, J. Jonasson, D. Killian, J. Bangoh, M. Kachule, and K. Palamountain. This work appeared in the *Journal of Medical Internet Research* [108].

1.2 Sample Transport Optimization

Malawi's national diagnostic network consists of 10 molecular laboratories, 27 district hubs, and over 700 healthcare facilities. Motorcycle couriers managed by Riders for Health Malawi are responsible for moving samples from healthcare facilities to district hubs, where they undergo administrative processing, and then onwards to molecular laboratories for testing. Couriers also transport batches of test results from laboratories to district hubs and healthcare facilities. Approximately 95% of samples transported within this network are associated with HIV viral load (VL) monitoring.

In Chapter 3, we propose a mixed-integer optimization model to generate efficient routes for couriers within the centralized diagnostic network. The model uses real-time data obtained from the information sharing system described in Chapter 2 to assess the demand for transportation at each location in the network on a daily basis, and selects routes to minimize transportation delays, subject to operational and cost constraints. We formulate this optimization problem as a Multi-stage Dynamic Vehicle Routing Problem (M-DMVRP) and implement our solution in a rolling horizon framework that leverages historical sample volume data to forecast the expected demand for transportation in subsequent days.

We assess the performance of the optimized transportation routes in a field trial conducted in three districts in Malawi and demonstrate that this approach offers significant advantages over fixed weekly transportation schedules. Key results include a 55% reduction in the number of unnecessary stops made by couriers at locations with no demand for transportation, as well as a 25% decrease in the average delays associated with sample transportation.

The success of this work has significant implications for other diagnostic networks in the region which rely on similar transportation systems. Our approach provides a roadmap for increasing the responsiveness and flexibility of ST systems in order to improve the efficiency and accessibility of diagnostic testing. The methods developed in this study may also be relevant to other logistical challenges within resource-limited healthcare systems, including the strengthening of supply chains for vaccines, drugs, and other medical supplies.

Joint work with S. Deo, J. Jonasson, M. Kachule, and K. Palamountain. This work is under review for publication in *Manufacturing & Service Operations Management* [78] and was awarded the 2020 INFORMS *Doing Good with Good OR* prize, the 2021 MSOM *Practice-based Research* award, and received second place in the 2021 POMS *College of Healthcare Operations Management Best Paper competition.*

1.3 Optimal Deployment of Point-Of-Care Technology for Viral Load Monitoring

New technologies for *Point-of-Care* (POC) testing enable clinicians to perform sophisticated diagnostic tests within healthcare facilities where patients are treated, returning results within a time frame of 1-2 hours. POC testing has the potential to provide faster, more reliable VL monitoring in resource-limited settings, where centralized diagnostic tests are often subject to substantial delays. However, POC testing can also be significantly more expensive than centralized testing, especially in facilities with low volumes of diagnostic samples.

In Chapter 4 we conduct a comprehensive analysis of the cost and impact of implementing POC VL monitoring within the existing centralized diagnostic network in Malawi. We develop a mixed-integer model to determine the optimal allocation of POC instruments to healthcare facilities under a variety of operational constraints, and evaluate several strategies to improve the cost-effectiveness of POC testing.

Our results demonstrate that transitioning approximately 20% of national sample volumes to POC testing could have a significant impact on health outcomes for people

living with HIV, particularly if these tests are appropriately targeted at high-risk patients who are more likely to benefit from faster clinical follow-up. In order to maximize the number of high-risk patients that can be reached, POC instruments at larger healthcare facilities should be used to provide rapid near-POC testing for samples collected at smaller facilities which have insufficient demand to justify the cost of an onsite POC instrument.
Joint work with S. Deo and J. Jonasson. This work is in preparation for journal submission.

1.4 Globally Optimized Survival Trees

Survival analysis is a cornerstone of healthcare research and is widely used in the analysis of clinical trials as well as large-scale medical datasets, Electronic Health Records, and insurance claims. A key advantage of survival analysis models is their ability to account for incomplete observations or *censored data* in which the outcome of interest is generally the time to an event (e.g., death, onset of disease), but the exact time of the event is unknown for some individuals (e.g., patients lost to follow-up in longitudinal studies).

Chapter 5 focuses on decision tree models for survival analysis. Tree-based models are increasingly popular due to their ability to identify complex relationships that are beyond the scope of parametric models, as well as their interpretable structure. Most decision tree algorithms generate models by iteratively adding partitions in a greedy fashion, which can result in unnecessarily complex models and poor out-of-sample performance. We present a new Globally Optimized Survival Trees (GOST) algorithm that leverages mixed-integer optimization (MIO) and local search techniques to generate globally optimized survival tree models.

We provide a detailed discussion of several accuracy metrics for survival models and use these metrics in a comprehensive analysis of the performance of the GOST algorithm. Our results demonstrate that the GOST algorithm improves on the accuracy of existing survival tree methods in both simulated and real-world datasets.

We illustrate how the algorithm can be applied to predict the risk of adverse events associated with cardiovascular health in the Framingham Heart Study (FHS) dataset.

Joint work with D. Bertsimas, J. Dunn, and A. Orfanoudaki. This work appeared in *Machine Learning* [23].

Chapter 2

Design and Implementation of a USSD System for Tracking Daily Sample Volumes in Malawi's Diagnostic Network

2.1 Introduction

Many populations in sub-Saharan Africa rely on rural health clinics as their primary point of entry to a broader healthcare system. These health clinics often lack the staff and equipment to conduct diagnostic testing onsite, but refer patient samples to a relatively small number of centralized laboratories capable of diagnostic analysis [113]. The effectiveness of programs targeting active diseases in the region, such as HIV-AIDS, Tuberculosis (TB), and Malaria, is therefore closely related to the performance of sample transportation (ST) systems, which transport patient samples and test results across the difficult terrain separating health facilities and molecular laboratories [191].

Sample transportation systems in many sub-Saharan African countries operate without accurate information regarding the quantity and location of the patient samples and test results requiring transportation [109]. Consequently, these ST systems operate in push mode, where couriers visit facilities on fixed weekly or bi-weekly schedules [109, 28]. A common result of this operating mode is empty trips: courier visits to facilities where nothing was delivered to or transported from the facility. For example, an analysis of archival 2017-2018 courier data in Malawi revealed that 31% of courier visits to clinics were empty trips. This not only results in inefficient utilization of limited resources but also contributes to delays in receiving critical test results [99], which in turn leads to poor health-seeking behavior among the population [145] and can contribute to increased mortality rates [105].

An alternative to these ST push systems is a "pull system" in which couriers only visit facilities when patient samples or test results are ready for transportation to or from that location, thereby limiting empty trips. However, this type of ST system would require a reliable method to track the number of patient samples and test results requiring transportation across the diagnostic network. We hypothesize that these logistics data — specifically, the location and quantity of patient samples or test results ready for transport — can be collected with a low-cost information sharing system and can be used to create a ST system which is responsive to real-time needs.

In this study, we investigated the feasibility, adoption, and accuracy of a system leveraging a communications protocol that is standard on all mobile phones, known as Unstructured Supplementary Service Data (USSD) technology [163]. To conduct this investigation, we designed and developed a USSD data collection system (hereafter, the USSD system) to enable healthcare facilities in a diagnostic network to report daily sample volume data. We conducted a year-long field trial of the USSD system in Malawi from July 2019 to July 2020 to determine if our system would enable the timely collection of accurate information.

2.2 Intervention Development and Design

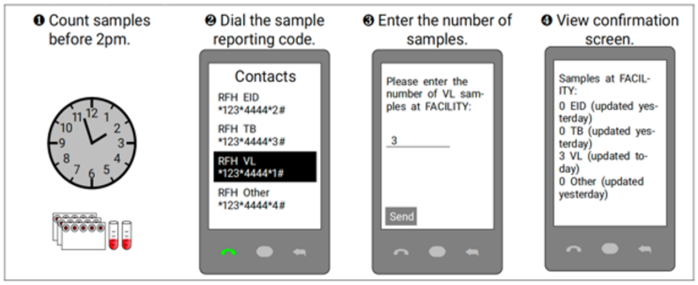
The critical design questions we faced when developing the information-sharing system were:

- What technology would health workers use to submit sample volume reports?
- How would those reports would be structured?

Given the positive findings regarding the feasibility of mHealth initiatives in low- and middle-income countries [3] and the rapid growth in mobile network coverage in sub-Saharan Africa over the past decade [9, 88], we elected to design our system so that health workers could submit reports with a mobile phone. As less than half of the mobile phones in the region are smartphones [88], we based our sample volume data collection system on USSD technology – a mobile communication protocol accessible to both smartphones and basic feature phones.

In a USSD system, users are provided with a numeric code that the users dial to access a structured menu of options. With an appropriate series of key presses using the mobile phone's keypad, users can navigate through the options menu and submit information, similar to how a short message service (SMS) system conveys information through text. One advantage of USSD over SMS, especially in the context of data collection, is that USSD's built-in menu system allows for structured responses, leading to built-in data validation [163]. This also reduces the amount of effort required to leverage the data in a systematic way, which is crucial to managing the growing volume of healthcare data received through mobile phones [148].

In our USSD system, a specific USSD reporting code is assigned to every health facility operating within a diagnostic network, and each type of diagnostic test offered in the network is assigned a unique numeric designator. To report the number of patient samples for a specific type of test, a designated health worker at a designated facility can use any mobile phone to dial that facility's USSD reporting code and the numeric designator for the diagnostic test to connect to the USSD system. Once connected, they can report the number of patient samples via a text-based interface. Upon entering all the required information, the user receives a confirmation text message containing a summary of the submitted report. See Figure 2.1 for more details.


❶ Count samples before 2pm.
12
1
2
3
4
5
6
7
8
9
10
11
❷ Dial the sample reporting code.
❸ Enter the number of samples.
❹ View confirmation screen.
Send

Contacts
RFH EID
*123*4444*2#
RFH TB
*123*4444*3#
RFH VL
*123*4444*1#
RFH Other
*123*4444*4#

Please enter the number of VL samples at FACILITY:
3

Samples at FACILITY:
0 EID (updated yesterday)
0 TB (updated yesterday)
3 VL (updated today)
0 Other (updated yesterday)

1. A health worker counts the number of samples, by sample type, that are prepared for transport.
2. The health worker dials a USSD code specifically assigned for the health worker's clinic and the type of sample (HIV early infant diagnosis (EID), HIV viral load monitoring (VLM), Tuberculosis (TB), Other) which has been saved on the health worker's phone and is displayed on a poster in the health facility.
3. The health worker enters the number of samples currently awaiting transportation at the facility.
4. The health worker views a confirmation screen which summarizes how many samples of each type have been reported for the facility.

Figure 2.1: USSD system reporting instructions.

2.3 Methods

2.3.1 Study Setting

As of 2018, 9.2% of adults (15–49 years) in Malawi were living with HIV [216] and the country's TB incidence rate was 181 per 100,000 people [231]. The Malawi Ministry of Health (MOH) operates a diagnostic network of approximately 700 widely-distributed community health clinics, 27 centrally located district health offices (DHOs), and 10 regionally-aligned molecular laboratories. The structure of the Malawi diagnostic network is representative of the diagnostic networks of many other countries in the region [191].

Since 2016, the Malawi branch of the nonprofit organization Riders 4 Health International (R4H) has managed the transportation of diagnostic samples for HIV viral load (VL) monitoring, HIV early infant diagnosis (EID), and TB testing. R4H maintains a team of over 80 motorcycle couriers based at district transportation hubs (usually located at DHOs) who visit health facilities and laboratories according to fixed weekly schedules.

In collaboration with MOH and R4H, we identified three districts in Malawi to test the USSD system: Salima in the Central Region, Rumphi in the Northern Region, and Phalombe in the Southern Region. The diagnostic networks in these districts each contained between 15 and 18 facilities and relied on a molecular laboratory outside of the district to conduct diagnostic testing, making these districts a representative sample of R4H's typical ST operations in rural and semi-rural areas.

This field trial was approved by Malawi's National Health Sciences Research Council. Evaluation by the MIT Committee on the Use of Humans as Experimental Subjects determined the trial did not constitute human subjects research as defined in Federal Regulations 45CFR46.

2.3.2 System Implementation

In early 2019, we contracted with a local vendor in Malawi to develop the user interface and information technology infrastructure. The vendor also managed the daily operation of the system, which included storing incoming data, providing an SMS gateway to send reminder messages to users when appropriate, and contracting with the cellular network providers to enable free provision of the USSD service to health workers.

In May 2019, we asked the facility in-charge from each of the participating health facilities to nominate one, two, or three staff members with personal mobile phones to enter data. In June 2019, we conducted three 2-hour training sessions (1 per study district) to train 150 health workers to use the USSD system. During the training sessions, we (i) introduced the USSD system to the participants, (ii) taught study participants how to access the system and submit reports using their personal mobile device, and (iii) provided reference posters and flyers reminding participants how to access the system for participants to display at their facilities. A field team, consisting of a local field manager and three local research assistants, monitored the implementation and addressed technical and logistical challenges through regular communication with health workers, district lab technicians, and R4H couriers via phone calls and text messages.

The USSD system was officially launched in the study districts in July 2019. Health workers were asked to report the number of patient samples waiting to be transported at the end of each day. Facilities were expected to submit a report every day, even if they had not collected any samples or prepared any new samples for transportation. Health workers and the field team were sent the following series of automated daily reminder messages to increase participation.

- 8:00 am - Health workers at each facility are sent a message notifying them whether or not a courier will visit their facility later that day as well as a reminder to report sample volumes.
- Noon - Health workers at each facility are sent a second reminder to report sample volumes.
- 1:30 pm - Members of the field team are sent a summary of the facilities for which a report has or has not been submitted.
- 2:15 pm - Health workers at facilities missing all or part of a complete daily report are reminded to report sample volumes.
- 3:00 pm - Members of the field team are sent an updated summary of the facilities for which a report has or has not been submitted, and a comparison of each facility's current report with their previous report (as an unusual increase/decrease from the previous report may indicate that the facility is reporting incorrectly).
- 4:15 pm - Members of the field team are sent a notification informing them whether couriers submitted reports to the courier database. Reports submitted to the courier database are summarized by facility, compared to that facility's most recent USSD report, and sent to members of the field team to assess facility reporting accuracy.
- 7:00 pm - Members of the field team are sent an updated list of the couriers who have or have not submitted reports to the courier database. A summary of reports submitted to the courier database by facility are

recompiled to capture any updates, re-compared to that facility's most recent USSD report, and sent to members of the field team to assess facility reporting accuracy.

Based on notifications received from health facilities, the field team sought out any unusual participation patterns such as intermittent, erratic, and/or extended periods of no participation. Upon detection of unusual participation patterns, the field team was authorized to address these patterns directly with the participant. In situations where an ordinarily reliable participant simply forgot to report, or health workers in the same facility failed to properly delegate reporting responsibilities, the field team could contact the designated staff member at the facility to remind them to submit their daily report or to delegate reporting responsibilities to a different health worker when the primary contact was not at the facility. If the field team identified intermittent network coverage as the cause of a missed report, the field team could delegate reporting responsibility to someone with a network connection. In addition, the field team could also ask the courier to hand-deliver a message to the responsible health worker, to identify someone else at the facility to accept reporting responsibilities, or to request escalation to a higher authority at the non-reporting facility in email notifications regarding their facility's participation.

The field team also monitored accuracy of reports and intervened directly with participants if they observed a pattern of low accuracy reports. As in the case of poor participation, the field team could use combination of phone calls, text messages, and hand-delivered messages to identify the root causes of data inaccuracies and address them. The preferred approach for improving accuracy of reports was to provide additional instructions to the non-compliant participant. If a training update failed to address the situation, more drastic measures (e.g., requesting transfer of reporting to another staff member) could be adopted.

2.3.3 Evaluation Framework

As part of the USSD system implementation plan, we elected to evaluate the feasibility, adoption, and accuracy of the USSD system using relevant descriptive statistics. Feasibility and adoption are common evaluation domains in intervention assessment literature [177, 165]. Accuracy, while not a common evaluation domain regarding health interventions, is relevant in the context of mobile-device based data collection systems [159]. Table 2.1 lists the guiding questions and associated metrics for assessing system performance within the three domains. The feasibility of the USSD system depended on whether each facility had access to the technology required to participate in the system. Therefore, we identified the number of facilities employing someone with a mobile device who was willing to participate in the study and the number of facilities in the field trial districts receiving service from a wireless network provider.

To assess adoption of the USSD system, we monitored specific participation-related metrics: percent of facilities reporting by day, individual facility participation over the course of the field trial, and the longest period each facility went without participating.

We determined the accuracy of the USSD system by comparing submitted USSD reports to program data. A data report was deemed accurate if the reported number of patient samples of a given type ready for delivery and the actual number of patient samples ready for delivery, as determined by the courier database, were identical.

2.3.4 Data

We used data from four distinct sources to calculate the metrics listed in Table 2.1: the USSD system database, the courier database, a survey administered to members of the field team, and the attendance roster from the USSD system training sessions.

Every data report submitted through the USSD system over the field trial was archived in the USSD system database. This database included the facility name, the date and time, the user's mobile device number, the sample type, and the number of patient samples reported by the user for every data report.

The courier database contained sample-specific information submitted by the

Table 2.1: Evaluation Framework

Domain	Guiding Question	Metric
Feasibility	Do facilities have access to mobile devices?	The fraction of facilities for which a personal mobile phone was registered.
	Do facilities receive a mobile network signal?	The fraction of facilities where insufficient network connection never prevented that facility from submitting a report.
		The fraction of facilities where staff members at that facility submitted a daily report for at least seven consecutive days.
Adoption	Are facilities participating?	The fraction of facilities that reported/failed to report each day by sample type.
		The fraction of total reporting days over the trial period when each facility reported/failed to report.
		The largest number of consecutive days a facility failed to report.
	Which operational factors influenced facility participation?	The fraction of facilities where an insufficient understanding of the USSD system on the part of health workers prevented USSD participation.
		The fraction of facilities where hardware limitations prevented USSD participation.
		The fraction of facilities where health worker work-load prevented USSD participation.
		The fraction of facilities where health worker absences prevented USSD participation.
		The fraction of facilities where health worker forgetfulness prevented USSD participation.
Accuracy	How accurate is the data reported by participating facilities?	The average and variance of the difference between reported and actual sample volumes.

courier upon completion of the courier's daily route to a data collection system operated by R4H. The sample data captured in the courier database consisted of the sample's identification code, the name of the facility the sample originated from, the date of sample collection, and the date of sample pick-up from the originating facility.

Upon completion of the study, we administered a survey to the research assistants to assess barriers to system participation. For each of the 51 facilities included in the field trial, the research assistants were asked to answer the following inquiries:

1. Estimate the number of times a particular event, including poor network reception, caused each facility in the research assistant's district to fail to report.

2. Rate the effectiveness of the following techniques on participation by facilities in their district:

    - SMS messages.
    - individual messages via a popular internet messaging platform.
    - phone calls.
    - asking a courier to deliver a message.
    - in-person facility visits.
    - group messages via a popular internet messaging platform.

We compiled a master attendance roster by combining the individual attendance rosters that were recorded at each of the three USSD system training sessions. These rosters included the name of each training participant, the facility the participant represented, the participant's staff position, and the participant's contact information.

2.3.5 Analysis

To calculate the percent of facilities for which a personal mobile phone was registered, we reviewed the master training attendance roster. Attendance at a training event by an employee from a given facility indicated that the employee owned a mobile phone and was willing to use their device to submit data reports to the USSD system. To calculate the percent of facilities with a sufficient network connection, we summarized the survey responses regarding the frequency with which poor network connectivity prevented each facility from participating.

To determine the number of facilities for which a USSD report was submitted for at least seven consecutive days, we analyzed the facility name and date of every report submitted to the USSD system database. Aggregating and/or summarizing data in the USSD system database also allowed us to measure all three metrics associated with the first guiding question in the Acceptability Domain in Table 2.1.

We calculated the accuracy of the reported data by comparing the reported data in the USSD system database to the courier reports in the courier database, which captured the number of patient samples collected from the healthcare facilities.

All data analysis was conducted with R (v4.0.0) and RStudio (v1.2.5042). Reported p-values were calculated using one-sided non-parametric Mann-Whitney tests (unless otherwise noted).

2.4 Results

2.4.1 Descriptive Statistics

Over the study duration (July 2019 - July 2020), participating facilities submitted 37,771 reports to the USSD system, accounting for 48,852 patient samples. The majority of these patient samples (83.8%) were VL samples (n = 40,952), while 6.1% were EID samples (2,979), 5.9% were TB samples (2,859), and 4.2% were classified as "Other" (2,056). Of the samples reported, 43.7% (21,355) originated in Phalombe, 35.1% (17,155) originated in Salima, and 21.1% (10,342) originated in Rumphi. Table A.1 contains sample volume statistics by facility.

2.4.2 Feasibility

All participating facilities employed at least one individual willing to submit reports to the USSD system with a personal mobile device. Research assistants reported that an insufficient network connection never prevented 47% of the participating facilities from submitting a report to the USSD system, caused occasional submission problems in 24% of facilities, and caused frequent problems in 29% of facilities.* Our analysis of the USSD system database data also revealed that each facility had at least one seven-day period where the facility submitted a report every day.

2.4.3 Adoption

Figure 2.2 illustrates the daily sample reporting rates and the 7-day moving average of daily sample reporting rates for the three patient samples types between July 2019 and July 2020. At the beginning of the study, only 10-20% of facilities participated each day. However, after three weeks, daily participation rates rose and remained between 53% and 98% for VL monitoring, between 51% and 98% for EID samples, and between 43% and 96% for TB samples.

Between August 2019 and January 2020, the average participate rate increased gradually for VL (63% to 87%), EID (60% to 85%) and TB (54% to 79%) samples with notable but temporary declines during the second half of November and December (due to a combination of mobile network outages and facility closures for the holiday season). For the final six months of the trial (February 2020 to July 2020), the average participation rate remained at or above 75% for all three sample types. This is particularly notable given the disruptions associated with the Covid-19 pandemic over this period.

The distribution of participation rates is shown in Figure 2.3, where the facility participation rate is calculated as the percentage of days on which a facility reported out of the total possible reporting days. On average, facilities provided a report 79% on of days (198 days out of the total 251 possible reporting days, $\sigma = 32.6$ days). The median number of days a facility reported was 204 days (81%), with a range from 121 days (48%) up to 245 days (98%). Facilities in Phalombe reported less frequently

*Feedback from research assistants noted that network issues were more common in facilities located near Malawi's borders, and that many staff at these facilities reported using SIM cards associated with cellular networks in neighboring countries. The USSD system was only accessible on Malawi's cellular networks.

Figure 2.2: Facility participation by sample type, shown as daily participation percentages and as a 7-day moving average.

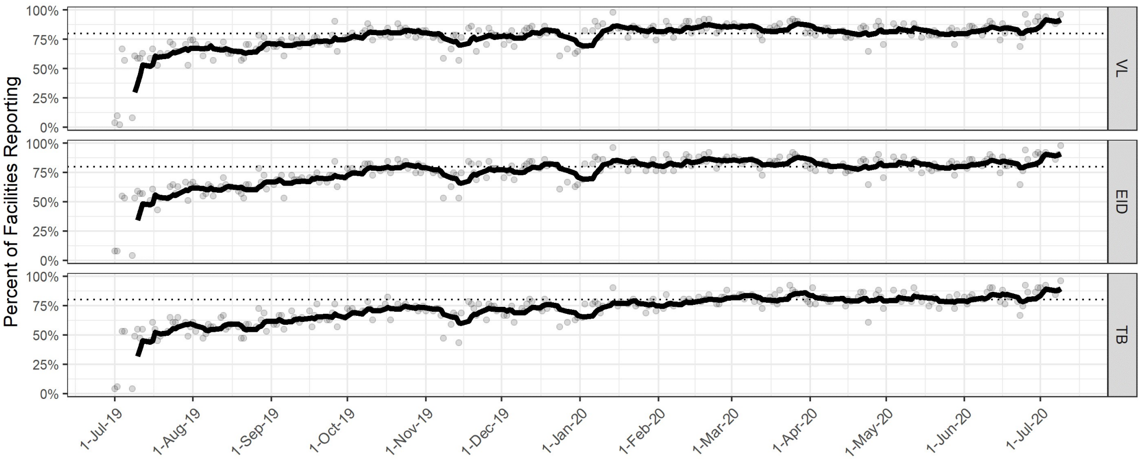
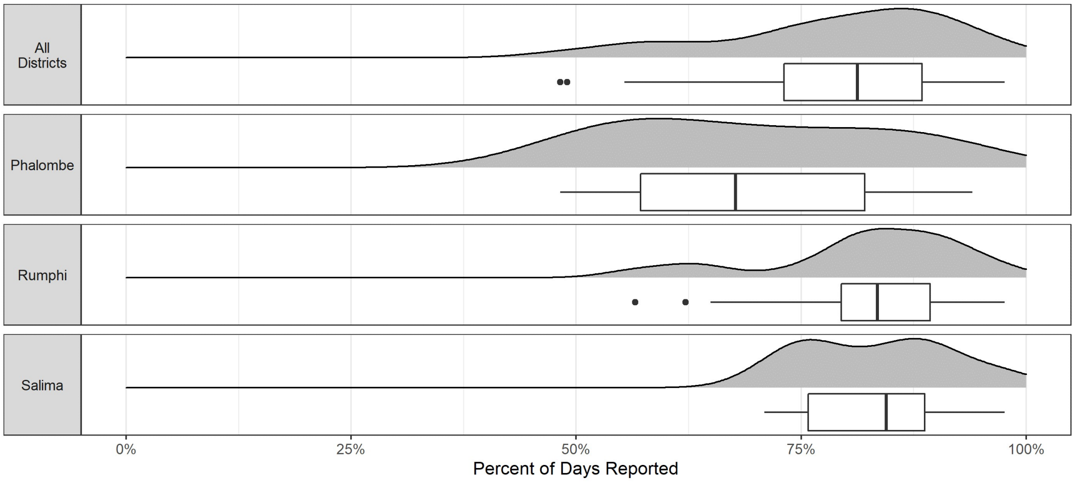


Figure 2.3: Distribution of facility reporting frequency.



than facilities in Salima (P = .003) and Rumphi (P =.01) on average.

Of the 51 health facilities, 45% of facilities (n = 23) never went more than one business week (5 days) without submitting a report, and the longest any facility went without providing a report was 30 days. On average, the longest period a facility went without submitting a report was 8.65 days ($\sigma = 6.33$).

Table 2.2 shows the frequency of participation challenges faced by health facilities according to the three local research assistants. Each cell shows the number and percent of facilities reported to experience a specific concern to the given extent.

Table 2.2: Survey results on the causes of reporting issues at each of the 51 participating facilities.

Reason for No Report	No Problems	Occasional Problems	Frequent Problems
Insufficient Mobile Network Reception	24 (47%)	12 (24%)	15 (29%)
Phone Issues (e.g., low battery, broken phone)	29 (57%)	18 (35%)	4 (8%)
Staff are Absent	20 (39%)	21 (41%)	10 (20%)
Staff are Too Busy	26 (59%)	15 (29%)	6 (12%)
Staff Do Not Understand How to Use the System	43 (84%)	8 (16%)	0 (0%)
Staff Forgot to Report	4 (8%)	39 (76%)	8 (16%)

Recurring compliance issues were most often due to poor network reception, while the occasional non-compliance issue was most likely due to forgetfulness on the part of the health worker. Staff absence at participating facilities also caused reporting issues at a majority (60%) of the participating facilities. These research assistant survey results also suggest that health workers had an adequate understanding of the system and that poor staff training did not cause any reporting problems in over 80% of the facilities.

2.4.4 Accuracy

Figure 2.4 illustrates the daily percentage accuracy of reports for each sample type, and the corresponding 7-day moving averages. The daily accuracy for VL reports slowly improved over the first two months of the field trial and settled at around 80% for the remainder of the trial. Unlike VL reports, the daily accuracy of EID and TB reports did not change substantially throughout the trial, with the daily accuracy of VL reports exhibiting greater variance ($\sigma = 7.23$) than both EID daily reporting accuracy ($\sigma = 5.25$; $P < .001$; Levene's Test) and TB daily reporting accuracy ($\sigma = 5.38$; $P < .001$; Levene's Test) . On average, 81% of the daily VL reports, 89.2% of the daily EID reports, and 88.2% of the daily TB reports were accurate.

The distribution of data accuracy by sample type across facilities is displayed in Figure 2.5. The accuracy of EID reports exhibited the least variation ($\sigma = 0.01$), followed by VL reports ($\sigma = 0.11$), and then TB reports ($\sigma = 0.14$). Median facility reporting accuracy were 82% for VL, which was lower than that for EID (91%; P =

Figure 2.4: The number of accurate reports by sample type, shown as daily accuracy percentages and as a 7-day moving average.

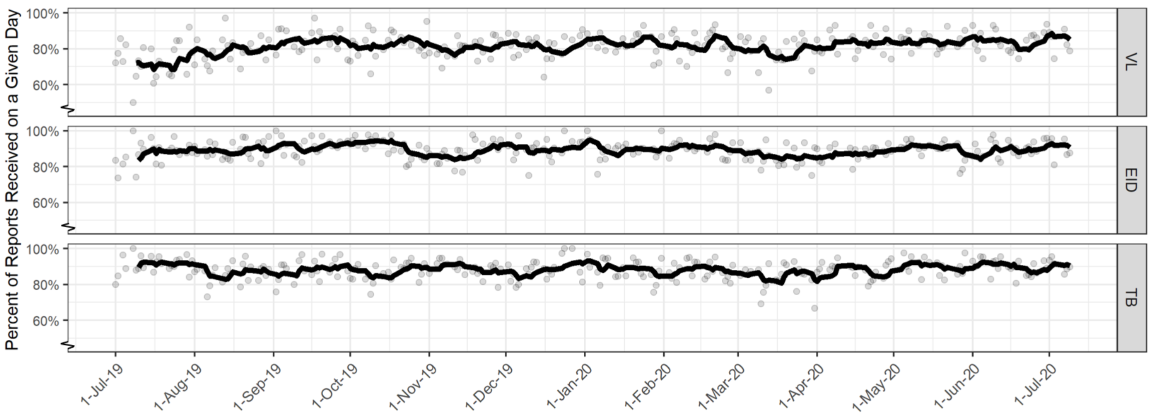
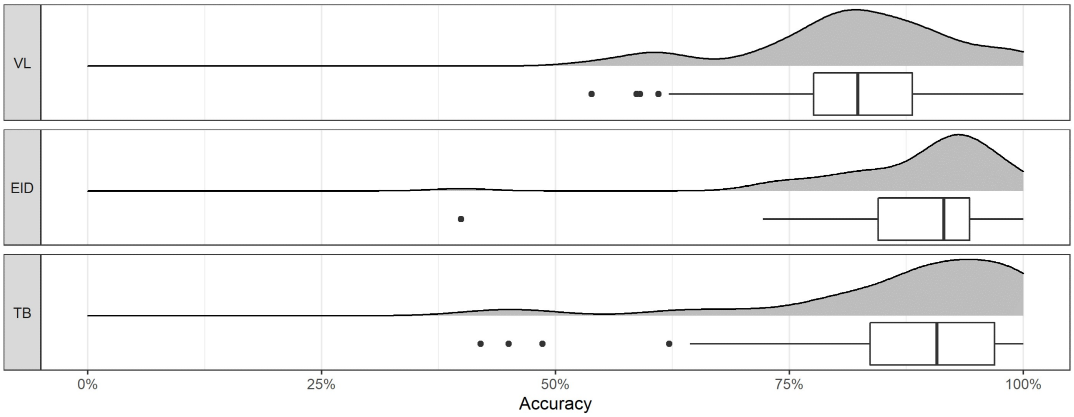


Figure 2.5: Distribution of facility reporting accuracy by sample type



.001; Paired Mann-Whitney) for TB (91%; P = .001; Paired Mann-Whitney). For each sample type, over half of the facilities submitted accurate reports on more than 80% of the days in the field trial.

2.5 Discussion

2.5.1 Summary of Findings

We designed a system whereby staff members in geographically dispersed healthcare facilities could report, via any mobile device, the number of diagnostic samples prepared for delivery to a diagnostic testing laboratory. Between July 2019 and July 2020, we conducted a field trial of the system in 3 districts in Malawi to assess the system's feasibility, adoption, and accuracy. The results from our field trial suggest that the USSD system is a feasible, adoptable, and accurate tool for assembling accurate daily reports on the quantity and location of transportation-ready patient samples.

2.5.1.1 Feasibility

The feasibility of the USSD system is driven by the ease with which facilities can submit reports using a mobile network connection and a mobile device. While mobile network coverage varies by country, our findings with respect to the number of facilities able to submit a report through the system illustrate the potential of using mHealth systems to link rural health facilities to central operations managers in sub-Saharan Africa, especially as mobile network coverage continues to improve across the region [88]. We also found at least one person at each facility with a mobile phone—a significant result, given that less than half of the population in the region owns a mobile phone [6]. While prior work on mHealth initiatives among the general population has, at times, found poor eligibility rates among potential participants driven by low mobile phone ownership in the region [12, 221], our findings suggest that mHealth efforts requiring ownership among health workers may be more feasible than those requiring ownership among the general population (see also [194]). Additionally, it is likely that the use of USSD in a region where smartphones constitute less than 40% of all mobile phones with lower-tier connections (i.e., 1G/2G as against 4G/5G) also contributed to system feasibility [88]. The use of health workers' personal phones avoided the additional cost of deploying and maintaining new devices in the field and leveraged familiarity with their devices to enhance system usability [19].

2.5.1.2 Adoption

Adoption of the USSD system improved consistently over the course of the study and peaked near 90% toward the end, with the exception of small and temporary declines coinciding with personnel transitions (i.e., staff reallocations and annual training sessions) and holiday seasons. These findings are comparable to results from similar mHealth studies conducted in the region [194, 19], despite our study requiring more frequent reporting than other studies and doing so without monetary compensation or other participation incentives. Based on the calculated descriptive statistics, we attribute the wide adoption of the USSD system to close collaboration with MOH representatives. This collaboration secured the support of senior government officials who encouraged participation by health workers at the facilities. Additionally, this collaboration improved the chances that that the system's design complemented health workers' existing responsibilities rather than adding to them, which is known to increase the likelihood of system adoption [12, 194, 107]. The efforts of the field team in their role as real-time participation monitors and problem-solvers also influenced the observed participation rate throughout the study.

2.5.1.3 Accuracy

The existence of the courier database, and our ability to access that data, played a significant role in ensuring high accuracy of records (greater than 80%). In contrast with prior mHealth initiatives [19, 132], the courier database allowed us to assess the accuracy of every report submitted through the USSD system with minimal delay, and to provide feedback, via the field team, to correct improper reporting behavior. Providing timely and relevant feedback to a health worker regarding their reporting behavior likely contributed to the overall reporting accuracy achieved in the field trial [91].

2.5.2 Study Strengths and Limitations

The scope of our study is limited to establishing the feasibility, adoption, and report accuracy of a USSD system for collecting information on the quantity and location of patient samples prepared for delivery in the diagnostic network. It is expected that this information is useful for avoiding unnecessary health facility visits, but the rigorous quantification of this effect on ST operations requires additional research (see Chapter 3). In addition, the results presented in Table 2.2 regarding the operational factors affecting facility participation are based on data collected indirectly through a survey administered to the three research assistants in the field. Ideally, this data should have been collected directly from each facility, as this would provide more granular information about the operational drivers for participation. However, collecting operational data on a daily basis was beyond the scope of this effort and the constant rotation of health workers into and out of facilities over the course of the study made it infeasible to conduct an end-of-study survey at each facility. We believe that the use of research assistants was the next best solution, as they were in regular contact with multiple health workers from each facility and were aware of their experiences with the USSD system. Regardless of this limitation, the main objective of the study was to assess the feasibility, adoption, and accuracy of collecting information using the USSD system, all of which can be evaluated using primary data sources.

While our study exhibits numerous strengths—including the fact that the structure of the system allowed us to determine the accuracy of every submitted report and a sustained field implementation for a year—a notable highlight of our study is that it demonstrates a novel use of mHealth technology to significantly improve information sharing in diagnostic networks, which have a similar structure in many in low- and middle-income countries. Previous mHealth studies have investigated how mHealth technology can improve healthcare delivery through the wide dissemination of health-related information [3], providing patient-specific reminders and/or results to patients [204, 112], connecting healthcare providers at different levels in the healthcare network [28, 119], and monitor medical supply stock levels [194, 19, 119], among other applications [12, 58]. This study, to the best of the authors' knowledge, represents the first application of mHealth technology track the location of samples in a large-scale diagnostic network.

2.5.3 Scalability

The USSD system is extremely scalable from a technological perspective, as there is no requirement to purchase a specific mobile phone, mobile phone airtime, or any other system-specific technology. Expanding the system to operate with new facilities and/or new diagnostic tests simply requires assigning USSD identification codes and training new users, which is a relatively simple process due to familiarity of mobile users in the region with USSD technology through other applications [163].

As explained earlier, the USSD system was designed to minimize any impact on the workload of health workers and avoid disrupting their routines, which should positively affect adoption in other health facilities [154, 195]. Further, health workers who currently use the USSD system can train new staff at their facilities to use the system or share their experiences when they are transferred to new facilities, speeding up adoption at other sites. Scalability may be adversely affected by continued reliance on the field team for data monitoring and supervision. As we develop scale-up plans in collaboration with R4H and MOH, we believe that incorporating these tasks into the roles of senior personnel within the ST systems, such as the regional coordinators who oversee ST operations within the districts, will help to overcome this problem and increase scalability.

2.6 Conclusion

Malawi's diagnostic network is representative of many diagnostic networks operated in sub-Saharan Africa, both in terms of the network's structure and the challenges is faces [191]. The results of our study suggest that a USSD-based system is a feasible, adoptable, and accurate solution to the challenge of untimely, inaccurate, or incomplete data within these diagnostic networks. The scalability of the USSD system, along with the promising results of our study, suggest that the implementation of such systems has the potential to improve data visibility in diagnostic networks in resource-limited settings.

2.7 Acknowledgements

This study was supported by Grand Challenges Award #OPP1182217 from the Bill and Melinda Gates Foundation.

The research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number U54EB027049. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Chapter 3

Sample Transport Optimization

3.1 Introduction

Access to diagnostic testing services is a critical element of public health systems in countries with high prevalence of communicable diseases such as the Human Immunodeficiency Virus (HIV) and Tuberculosis (TB). In resource-limited settings, diagnostic services are frequently provided through hierarchical networks comprised of health facilities that collect samples (e.g. blood, sputum, etc.) from patients, and centralized laboratories that process these samples. This strategy provides low-cost, high-volume testing while ensuring that patients in remote areas have access to diagnostic services within their own communities [164, 11].

Sample transport (ST) systems enable the physical movement of samples and test results between health facilities, district hubs, and laboratories within these multi-stage networks, and contribute significantly to the efficiency of diagnostic service delivery [117]. In many countries in sub-Saharan Africa, ST systems are centrally operated with fixed weekly or monthly transportation routes necessitated by a lack of real-time ST demand information from health facilities. Consequently, ST operators are unable to dynamically allocate transportation resources to adapt to daily variability in ST demand, leading to unnecessary travel (when health facilities do not have samples for collection) and unnecessary delays (when samples at facilities have to wait for the next scheduled visit).

In this chapter, we describe the design and implementation of a new *optimized sample transportation* (OST) system that addresses these challenges, with the ultimate goal of reducing result turnaround times in Malawi's diagnostic network. The OST system comprises two components: (i) a novel data sharing platform to monitor incoming sample volumes at healthcare facilities, as described in Chapter 2, and (ii) an optimization-based solution approach for generating daily transportation schedules and courier routes, in response to current demand at each site. To this end, we formulate a general multi-stage version of the dynamic multi-period vehicle routing problem (which we refer to as M-DMVRP and define in Section 3.4), develop an optimization-based solution approach (Section 3.5), and numerically evaluate the performance of this solution approach for scheduling and routing couriers on synthetic multi-stage transportation networks (Section 3.6). As part of our numerical experiments, we compare our solution approach to a lower bound on the optimal solution, which can be generated for small instances of the problem (Section 3.6.1). We also benchmark our method against other scalable solution methods on networks of a realistic size (Section 3.6.2). Finally, we conduct a sensitivity analysis to illustrate how the performance of dynamic scheduling policies depends on the accuracy of the information collected through the information sharing platform (Section 3.6.3).

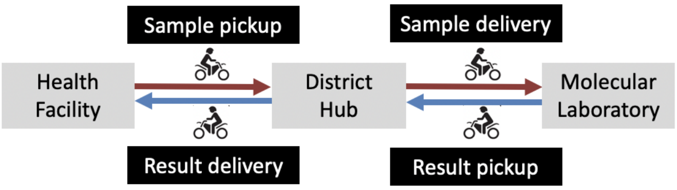
We implement and test this system in collaboration with Riders 4 Health Malawi (R4H), a non-profit organization that operates a national sample transportation system reaching approximately 700 healthcare facilities. We assess the impact of this approach in a pilot study conducted in three districts in Malawi (each comprising 15-18 health facilities served by two R4H couriers) from August–October 2019 (Section 3.7 provides details on the field trial design). Based on analysis of over 20,000 samples and results transported during this study, we show that the implementation of OST routes reduced average ST delays by approximately 25%. The ongoing implementation of this system also resulted in a 55% decrease in the proportion of unnecessary trips by ST couriers, demonstrating that optimized ST routes improve turnaround times without increasing unnecessary travel (Section 3.8).

From a practitioner standpoint, our work has implications for the operations management of diagnostic networks in other low- and middle income countries, particularly since difficulty in specimen transport has been reported as a key challenge to successful scale-up [117]. With significant scale-up efforts underway in Cameroon, Kenya, Namibia, South Africa, Swaziland, Thailand, and Uganda [188], our approach of combining a low-cost information sharing platform with an optimization-based scheduling and routing approach has the potential to inform the design of efficient diagnostic networks in each of those countries. Through the ongoing implementation of this system in Malawi, we demonstrate that our approach is feasible in a real-world setting and significantly improves the efficiency of ST operations. Furthermore, our model contributes to the existing literature on DMVRPs by introducing an interesting problem from an important healthcare setting which presents practically motivated challenges for transportation optimization. Specifically, the problem has multiple stages in which actions in one period (e.g., collecting samples) will generate demand for actions in subsequent period (e.g., moving those sample samples to the next stage). We present a novel solution approach to this problem—the combination of information collection to resolve short-term uncertainty and an optimization-based solution approach for scheduling and routing—which performs well in numerical experiments and through field implementation.

3.2 Background and Context

Malawi has one of the highest HIV prevalence rates in the world—approximately one million of its 20 million citizens are HIV positive [215]. Like many other countries in the region, Malawi's healthcare system has undergone significant changes in response the HIV epidemic, and Malawi has made significant progress in meeting the 90-90-90 goals, namely, that 90% of HIV positive citizens should know their status, 90% of these patients should be receiving long-term antiretroviral therapy (ART), and 90% of ART patients should be virally suppressed [214]. Large-scale diagnostic testing plays a crucial role in monitoring ART patients to ensure viral load suppression [217], and Malawi's National HIV Program has developed an extensive diagnostic network to

Figure 3.1: Diagnostic sample transportation network



support these needs. In 2019, this network processed approximately 640 000 viral load samples and 85 000 Early Infant Diagnosis (EID) tests.

3.2.1 Current Diagnostic Network

Malawi's diagnostic network consists of approximately 700 healthcare facilities, 27 district hubs, and 10 molecular laboratories. Patients access HIV testing and treatment at healthcare facilities, which range from small rural health centers to large regional hospitals. Samples collected at these facilities are transported to district hubs for administrative processing (essentially being entered into a database by a data clerk) and then forwarded to molecular laboratories for testing. After testing, results are printed at molecular laboratories and transported back to district hubs to be recorded and distributed to the corresponding healthcare facilities (see Figure 3.1). The turnaround time for this entire process generally takes several weeks, depending on the current load at molecular laboratories. In 2019, the median age of viral load samples on the day of testing ranged from 13 days to 38 days across different laboratories.

Since 2015, Riders 4 Health Malawi has operated a national ST network consisting of approximately 80 full-time motorcycle couriers based at district hubs around the country. These couriers operate on fixed weekly schedules that visit each healthcare facility on predetermined days (generally twice a week) to pick up samples and deliver results. Couriers in each district also make weekly trips to a molecular laboratory to deliver accumulated samples and pick up test results.

Although fixed weekly visit schedules provide regular and reliable sample transportation, this operational strategy is not necessarily efficient when there is a high level of variability in sample volumes. In 2017, approximately 30% of scheduled trips to healthcare facilities were unnecessary (i.e., there were no samples or results to be transported to or from the site), resulting in approximately 90 000 km of unnecessary travel. Poor data visibility is a significant barrier to addressing these operational inefficiencies, as most healthcare facilities have no formal communications infrastructure (telephones, fax, etc.). Frequent unnecessary trips increase the cost-per-sample of ST operations and also divert ST resources that could be used to reduce delays in other parts of the system.

3.2.2 Outline of Proposed System Changes

Our proposed system changes are based on two conjectures about the underlying causes of inefficiency in the status quo. First, the predetermined schedule, which ensures a weekly visit to every single clinic in the country, results in clinic-visits where no samples are picked up and no results are brought back. Such unnecessary trips are costly and negatively impact TATs through inefficient allocation of resources, which in turn has negative implications for public health through delayed or diminished treatment initiation. Second, a predetermined schedule for ST was adopted because of lack of real-time information about ST demand and lack of the computational skills and infrastructure required for designing optimal routes on a daily basis.

Our *optimized sample transportation* (OST) system therefore has two components. For information sharing, we leverage the USSD platform described in Chapter 2, which allows health workers at facilities within the ST network to communicate with the district office managing ST operations on a daily basis. This will ensure that ST operators have up-to-date information on the need for sample transportation in any clinic in their district. In Section 3.4, we formulate the decision problem of how to schedule and route couriers on a daily basis, assuming that the decision maker has access to the current ST demand information through the information sharing platform but must plan for stochastic arrivals on subsequent days.

3.3 Literature Review

In this section we review three streams of literature that are related to our work. We begin with a review of relevant models in the transportation literature in Section 3.3.1. We then provide an overview of the public health literature on diagnostic networks and ST systems in Section 3.3.2 and conclude by summarizing our contributions to the global health operations management literature in Section 3.3.3.

3.3.1 Vehicle Routing and Dispatch Problems

Our work is related to a range of extensions of the classical vehicle routing problem (VRP, see [83] for an overview). A stream of work on the dynamic pickup and delivery problem is concerned with scheduling the transportation of objects or people from an origin to a destination (see [21] for a review) but usually does not consider a multi-period horizon [228]. The dispatching literature considers the dynamic [e.g., 110] or deterministic [e.g., 13, 42] dispatching of orders to customers, usually focused on meeting a deadline for each order.

Our work belongs to the broad category of dynamic vehicle routing problems under uncertainty ([176] provide a review), which has been addressed using fixed schedules with recourse options [39, 50]. More specifically, the key uncertainty in our setting is on the demand side [see e.g., 7, 25, 84, 193, for single-period formulations of the VRP with uncertain demands]. Dynamic VRPs with stochastic demand have only recently been extended to multiple periods [211, 212], motivated by retail or field technical support settings. These models consider two classes of demand (early requests that must be served within the planning period or late requests that can be postponed to become early requests for subsequent periods). This literature assumes that demand that remains unmet for a multiple periods is lost, which simplifies the state space and eases the computational burden significantly. This is distinct from our setting, where samples and results remain in the system until they have progressed through each stage of the transportation network.

The emerging literature on the dynamic multi-period vehicle routing problem

(DMVRP, also referred to as the Tactical Planning VRP in [15]) is most related to our work. The DMVRP, first described by [30] and [227], comprises a sequence of vehicle routing problems that unfold over multiple consecutive periods. New requests for transportation are revealed at the start of each day, and these requests are either included in vehicle routes for the current day or postponed to be fulfilled on a subsequent day in combination with future requests. The objective of most DMVRP formulations is to minimize logistical costs (e.g., a combination of the number of vehicles required and associated fuel costs) subject to meeting delivery (or pick-up) time targets. Since future requests for transportation are not known in advance, a key challenge in DMVRP formulations is to ensure that unfulfilled demand carried over to subsequent days will not lead to infeasible or inefficient routes in future. Subsequent DMVRP models have introduced extensions of this problem. [10] describe a formulation that includes probabilistic information about uncertain future demand but do not explicitly present a stochastic dynamic formulation of the decision problem. [201] present an intractable robust adaptive optimization formulation of the DMVRP problem and propose two tractable approximations to bound its objective.

A key distinction between our models and the prior literature is that we model a multi-stage network in which the same vehicle capacity is used for transportation between a pick-up point (the health facilities) and an aggregation point (the district hub) as well as between the aggregation point and final destination (molecular lab).[*] This sharing of transportation capacity between line haul routes and last mile routes is frequently the case in low-resource settings, for example for the transportation of vaccines or essential health supplies. This extension is important since existing DMVRP models assume that future demand for transportation is independent of routing decisions and across locations. In contrast, a substantial portion of future demand in our setting is directly linked to current demand and current routing decisions. For example, collecting samples from a healthcare facility today will result in future demand for transportation at the next stage of the ST network (from a

[*] In our setting there are a further two stages when one considers the transportation of results back from the molecular lab to the district hub and from the district hub back to the health facility.

district hub to a molecular laboratory) in subsequent days. Additionally, in line with the objectives of R4H, our objective is to minimize the total average delays across the entire ST system by reallocating fixed transportation capacity based on current sample/result volumes at each location. This is distinct from the vast majority of VRPs under uncertainty, where the objective is usually to meet a hard constraint of delivery deadlines. Finally, existing DMVRP formulations model transportation requests as binary (i.e., demand for transportation at each location is either present or absent on a given day). In the ST network, transportation demand at each location accumulates over time as new samples and results become available. This introduces a more complex trade-off in our model; in addition to considering the geographic proximity of locations with unmet demand, our formulation must also prioritize trips to different sites based on the magnitude of demand at each location and the ongoing accumulation of demand on subsequent days.

3.3.2 Diagnostic Networks and ST Operations

Diagnostic networks in sub-Saharan Africa have experienced increased workloads in recent years due to the scale-up of HIV viral load monitoring programs [214]. Assessments of these growing systems have highlighted long delays in testing and result delivery as a significant barrier to effective ART treatment management, and longer diagnostic turnaround times (TATs) have been linked to poorer health outcomes in patients [147, 225].

Inefficient sample transportation is acknowledged as a key challenge in achieving the desired VL monitoring coverage and shorter turnaround times in sub-Saharan Africa [117, 164, 166, 188]. A 2014 study in Malawi highlighted significant delays associated with the transfer of results between laboratories and healthcare facilities [189], while analysis of 2016 data from Malawi's Laboratory Information Management System (LIMS) identified delays in sample transportation to laboratories as a likely explanation for longer TATs for sample testing at rural sites [135]. Other studies have identified similar concerns in Zimbabwe, Uganda, and South Africa [226, 109, 200].

Efforts to improve ST systems have thus far focused on reaching remote and under-served areas; replacing fragmented, ad-hoc transportation arrangements with centralized systems [109, 69, 106]; and leveraging simple mobile communication systems to expedite delivery of results [59]. Several studies investigating the potential applications of new technologies such as unmanned aerial vehicles (drones) have concluded that the cost and complexity of these solutions compares poorly to motorcycle couriers [156, 234]. Due to the lack of infrastructure in rural areas, successful sample transportation systems generally favor low-cost, low-tech operations that prioritize reliability and consistency.

We contribute to this literature by developing and implementing a system that enables dynamic ST routing in diagnostic networks by combining low-cost technologies (to collect previously unavailable data) and advanced analytics (for decision support) to reduce delays and unnecessary travel. Since diagnostic networks across many countries in sub-Saharan Africa have a similar structure, our approach has the potential to be replicated across the region to realize similar benefits.

3.3.3 Global Health Operations Management

Our work contributes to a growing body of literature that focuses on developing insights and models that are applicable to healthcare delivery in resource-limited settings. Various streams of prior work have examined how the funding and supply uncertainty in such settings affect procurement, inventory management, and distribution of supplies to health facilities [76, 120, 144]; how best to structure incentives in drug supply chains and medical device distribution [205, 121, 238]; how to motivate treatment adherence among Tuberculosis patients [202, 33]; how to optimize ambulance locations in developing countries [32]; and how to strengthen supply chains and distribution networks for vaccines, medicines, and healthcare products [111, 158, 57].

Of particular relevance to our work, recent research has focused on the design and management of HIV diagnostic networks in resource-limited settings. Specifically, [60] evaluate the impact of point-of-care technology in reducing delays in diagnostic services; [99] propose models for optimizing laboratory capacity allocation to improve efficiency across large-scale diagnostic networks; and [198] propose a mixed-integer optimization model to construct transportation routes for a high-volume diagnostic network with deterministic demand, where each facility is visited on a daily basis.

Our work extends this literature in two ways. First, we develop a system for optimizing the daily management of ST operations, whereas the previous literature has treated the existing operational implementation of ST systems as an invariable component of diagnostic networks. To this end, we develop a novel system for simultaneous information collection and optimization for scheduling ST couriers, which requires formulating and solving a version of the DMVRP which is tailored to ST operations. Second, while some prior work has been practice-driven in nature and has calibrated models using programmatic data, our work takes the next step of implementing our system in the field to demonstrate operational feasibility and for impact evaluation.

3.4 The Multi-Stage Dynamic Multi-Period Vehicle Routing Problem

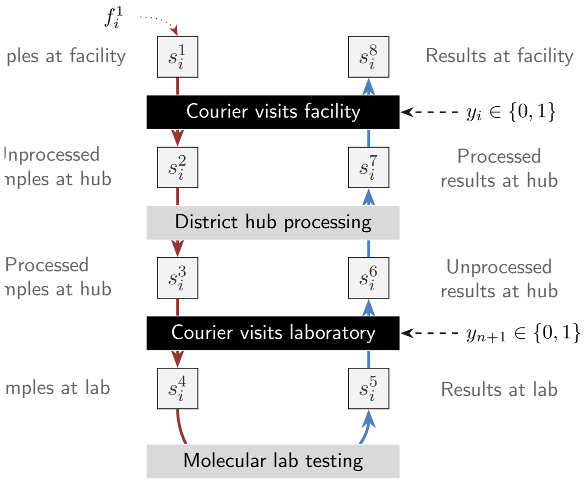
In this section, we define the Multi-Stage variant of the Dynamic Multi-Period Vehicle Routing Problem (M-DMVRP). This is the problem faced by ST operators when scheduling and routing ST couriers to visit health facilities (collecting samples and delivering results from the district hub) and the molecular laboratory (delivering samples from the district hub and collecting results from the laboratory).

3.4.1 Problem Description and Notation

Network Structure. We consider ST operations in a focal district. The diagnostic network of the district is defined by three types of locations (indexed by i); the district hub that serves as the depot for the ST operation ($i = 1$) in addition to collecting samples from patients, $n - 1$ additional health facilities which also collect samples ($i \in \{2, \ldots, n\}$), and the molecular lab ($i = n + 1$).

Decisions. On each day, the key decisions for the ST operator are which locations

Figure 3.2: ST network model.



to visit and how to route couriers to those destinations. We denote the daily courier routes by a binary routing matrix with entries $x_{i,j}(t)$, where $x_{i,j}(t) = 1$ if a courier route travels from location i to location j on day t, and $x_{i,j}(t) = 0$ otherwise. For convenience, we introduce auxiliary schedule variables $y_i(t) = \max_{1 \leq j \leq n+1} x_{i,j}(t)$ to indicate whether location i is visited on day t.

Sample Arrivals and Movements. Figure 3.2 illustrates the sequence of eight stages each sample moves through from the time that the patient visits the health facility until the results are returned to the health facility. In the first four stages, samples are collected at facilities, moved to the district hub, processed at the hub, and moved to the molecular laboratory for testing. In the last four stages, test results are printed at the molecular laboratory, moved to the district hub, processed at the hub, and delivered to the original healthcare facility.

On each day t, we denote the current sample/result volumes in the diagnostic network as an $n \times 8$ matrix $S(t)$, with each entry $s_{i,k}(t) \geq 0$ representing the total number of samples/results from facility $i \in \{1, \ldots, n\}$ currently in stage $k \in \{1, \ldots, 8\}$ of the diagnostic network. We denote the initial state of the system (at $t = 0$) by v_i^k. For subsequent periods, we define a set of variables $f_{i,k}(t)$ to describe the number of samples/results from facility i that enter stage k on day t;

$$s_{i,k}(t+1) = s_{i,k}(t) + f_{i,k}(t) - f_{i,k+1}(t). \tag{3.1}$$

The flow into stages 2, 4, 6, and 8 (see Figure 3.2 is completely determined by the visit schedule, since each time one of these locations is visited, the courier transports all available samples/results that can be moved to this location:

$$f_{i,k}(t) = y_i(t)s_{i,k-1}(t), \qquad \forall k = 2, 8, \tag{3.2}$$

$$f_{i,k}(t) = y_{n+1}(t)s_{i,k-1}(t), \qquad \forall k = 4, 6. \tag{3.3}$$

In contrast, the flow into the remaining stages (3, 5, and 7) is determined by the processing capacity at the district hub and the molecular lab (for processing samples, testing samples, and processing results, respectively).

$$f_{i,k}(t) = f_{i,k-1}(t - \omega_k) \qquad \forall k = 3, 5, 7, \tag{3.4}$$

where ω_k denotes the operational delays associated with the internal processes at these stages.

We denote the arrival pattern of new samples to each healthcare facility on each day by a random variable $u_i(t)$. The distribution of these random variables varies across facilities and depends on t, capturing the fact that many health facilities schedule HIV consultations on specific days of the week, resulting in periodicity in the arrival pattern (see numerical experiments in Section 3.6.1 for sensitivity to this periodicity). We assume that the distribution of $u_i(t)$ takes the form $P(u_i(t) = u) = g_{i,w(t)}(u)$, where $w(t) \in \{1, \ldots, 7\}$ indicates the weekday. As a result, the flow into the first stage of the ST system is

$$f_{i,1}(t) = u_i(t), \quad \forall t \in \{1, \ldots, T\}. \tag{3.5}$$

In summary, we represent the movement of samples through the system using the recursive function $Q(\cdot)$;

$$S(t+1) = Q(S(t), \mathbf{y}(t), \mathbf{u}(t)). \tag{3.6}$$

The multi-stage nature of our problem is captured in the network flow constraints (3.1)-(3.4). We note that this important feature of our setting, which requires the same transportation capacity to be used for transportation between various stages, has not been considered in the previous literature on DMVRPs (see Section 3.3).

Constraints. We assume that the number of couriers and vehicles is fixed and that variable costs (fuel and maintenance, in our setting) are directly proportional to the distance traveled by the couriers. The ST operator has a fixed long-term budget which can be apportioned to give an average travel budget of $\bar{d}$ km per day. We denote the daily distance traveled by the ST couriers (which is determined by the routing variables $x_{i,j}(t)$) by $d(t)$. The routes may exceed the daily distance budget on some days, as long as the average daily distances remain within the budget. We denote the cumulative budget surplus/deficit on day t as $b(t)$, where;

$$b(t+1) = b(t) + \bar{d} - d(t). \tag{3.7}$$

Objective. The ST operator's objective is to minimize the average turnaround time of samples the diagnostic network (i.e., days between sample arrival and result delivery), subject to an exogenous budget constraint which limits the number of feasible visits. Using Little's Law, the average TAT within the diagnostic network is proportional to the average number of samples/results within the system on a daily basis;

$$\Omega = \lim_{T\to\infty} \frac{1}{T} E[\sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{k=1}^{7} s_{i,k}(t)]. \tag{3.8}$$

3.4.2 Markov Decision Process Formulation

The sample transport problem described above can be represented as an infinite-dimensional MDP with state space $\mathcal{S}$;

$$\mathcal{S} = \{\Gamma = (S, w, b) | s_{i,k} \in \mathbb{Z}_+, w \in \{1, \ldots, 7\}, b^{min} \leq b(t) \leq b^{max}\} \tag{3.9}$$

and action space

$$\mathcal{A} = \{\mathbf{y} | y_i \in \{0, 1\} \ \forall i = 1 \ldots, n+1\}. \tag{3.10}$$

Based on the constraints of the ST system, we restrict the action space $\mathcal{A}$ to visit schedules that can feasibly be completed in a single day by the couriers available in each district. Provided that these constraints are met, we do not need to include the sequence of routes associated with each schedule, as the order in which facilities are visited has no impact on the system dynamics. We assume that the total daily distance traveled by all couriers, $d(t)$, corresponds to the distance of the shortest combination of feasible routes that visit that visit all of the locations required by the visit schedule $\mathbf{y}(t)$. We require that all feasible routing policies $\pi : \mathcal{S} \rightarrow \mathcal{A}$ do not exceed the budget deficit constraint $b^{min} \leq b(t)$. We include an empty visit schedule $\mathbf{y} = \mathbf{0}$ in our action space to ensure that at least one action is feasible from every state.

For simplicity, we assume that processing times at facilities and DHOs are equivalent to one day, and that the ST system operates 7 days a week. Both of these assumptions can be removed with appropriate modifications to the state space and action space.

For any action $\mathbf{y^1} \in \mathcal{A}$ and states $\Gamma^1 = (S^1, w^1, b^1), \Gamma^2 = (S^2, w^2, b^2) \in \mathcal{S}$, the transition probabilities are

$$P(\Gamma(t+1) = \Gamma^2 | \Gamma(t) = \Gamma^1, \mathbf{y}(t) = \mathbf{y}^1) = P(\mathbf{u}(t) \in \mathcal{U}_{S^1,S^2}) \times \tag{3.11}$$

$$P(w(t+1) = w^2 | w(t) = w^1) \times \tag{3.12}$$

$$P(b(t+1) = b^2 | b(t) = b^1, \mathbf{y}(t) = \mathbf{y}^1) \tag{3.13}$$

where where $\mathcal{U}_{S^1,S^2} = \{\mathbf{u} \in \mathbb{Z}_+^n | S^2 = Q(S^1, \mathbf{y}^1, \mathbf{u})\}$ is the (possibly empty) set of all vectors $\mathbf{u}$ that would satisfy the system dynamics equations for states S^1 , S^2, and the probabilities (3.12) and (3.13) are binary functions which indicate whether the distance budget and weekday of state Γ^2 are compatible with the current state Γ^1 and the distance of the route $\mathbf{y}^1$.

The objective of the MDP problem is to is to minimize the average TATs in equation 3.8. We therefore define the corresponding cost function

$$R(\Gamma(t), \mathbf{y}(\mathbf{t})) = \sum_{i=1}^{n} \sum_{k=1}^{7} s_{i,k}(t), \tag{3.14}$$

which depends only on the state variables $S(t)$. We then aim to find a policy, $\pi : \mathcal{S} \to \mathcal{A}$, that minimizes the value function

$$J_\pi(\Gamma(0)) = \lim_{T\to\infty} \frac{1}{T} E[\sum_{t=0}^{T-1} R(S(t), \pi(\Gamma(t))]. \tag{3.15}$$

The MDP problem is generally intractable, even for small networks, due to the size of the state space required to capture each of the different stages in the ST network. We therefore develop a more tractable optimization-based solution approach in section Section 3.5.

3.5 The OST Solution Approach to the M-DMVRP

We now describe the scalable and implementable *optimized sample transportation* (OST) solution approach to the decision problem presented in Section 3.4. We start with the deterministic equivalent of the problem, by deriving a network flow formulation for jointly optimizing the scheduling and routing of couriers (Section 3.5.1). We present a more computationally efficient variant for the case with an existing menu of feasible routes (Section 3.5.2) and then describe how that optimization model can be used as the basis for a solution approach to the stochastic problem (Section 3.5.3).

3.5.1 A Deterministic Scheduling and Routing MIP Formulation

Suppose that sample arrivals (at health facilities), sample registration times (at district hubs), and sample processing times (at molecular labs) are known with certainty at the beginning of the planning horizon. We now formulate a deterministic mixed integer optimization model (MIP) for a planning horizon of T days, to optimize the routing and scheduling of couriers in this setting.

Objective Function. The objective of our model is to minimize the average turnaround time of the diagnostic network (i.e., days between sample arrival and result delivery). As discussed in Section 3.4, this is proportional to the average sample volumes in stages $1 - 7$. However, since we optimize over a limited time horizon, we add a penalty for the number of samples that remain in the system at the end of the horizon[†], which results in the following objective;

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=1}^{7} s_{i,k}(t) + \sum_{i=1}^{n}\sum_{k=1}^{7} \rho_k s_{i,k}(T) \tag{3.16}$$

Distance Budget Constraints. We include three constraints to ensure that routes do not exceed the budget for long-term (over the planning horizon) or short-term (weekly) travel distance;

$$d(t) = \sum_{i}\sum_{j} D_{i,j} x_{i,j}(t), \quad \forall t, \tag{3.17}$$

$$\sum_{t} d(t) \leq \texttt{total\_dist}, \tag{3.18}$$

$$\sum_{t\in\{1,\ldots,7\}} d(t) \leq \texttt{weekly\_dist}, \tag{3.19}$$

where D is a matrix of pairwise distances between all locations, `total_dist` denotes the total distance that all vehicles can travel over the planning horizon, and `weekly_dist` denotes the total distance that all vehicles can travel in a single week.

[†] In our numerical experiments, the ρ_k coefficients in the penalty are set to represent the unavoidable delays for the samples remaining in the system beyond the planning horizon.

Routing Constraints. We include a set of standard VRP constraints to ensure the feasibility of routes. First, we require that the total number of routes does not exceed the number of available couriers ($\texttt{couriers}(t)$) on each day t, and that the schedule variables $y_i(t)$ reflect the routes $x_{i,j}(t)$;

$$y_i(t) \leq \sum_j x_{i,j}(t), \quad \forall t. \tag{3.20}$$

$$\sum_i x_{i,1}(t) = \sum_j x_{1,j}(t) \leq \texttt{couriers}(t), \quad \forall t. \tag{3.21}$$

Second, we require that locations other than the hub are visited at most once per day and that each location visited has one incoming and outgoing trip;

$$\sum_j x_{j,i}(t) = \sum_j x_{i,j}(t) \quad \forall i \in \{2, \dots, n+1\}, \quad \forall t. \tag{3.22}$$

Third, we ensure that trips to the molecular laboratory do not include any visits to health facilities;

$$\sum_{j \geq 2} x_{j,n+1}(t) = \sum_{j \geq 2} x_{n+1,j}(t) = 0, \quad \forall t. \tag{3.23}$$

Finally, we adapt the MTZ subtour elimination constraints proposed by [134] to ensure that each route begins and ends at the hub, and to limit the number of stops and the total distance of each route;

$$\begin{aligned}
&0 \leq h_i(t) \leq \texttt{max\_stops}(t) - 1 && \forall i \in \{2, \dots, n+1\},\ \forall t, && (3.24)\\
&h_j(t) - h_i(t) \geq 1 + (x_{i,j}(t) - 1) * \texttt{max\_stops}(t) && \forall i, j \in \{2, \dots, n+1\}\ \forall t, && (3.25)\\
&D_{i,1} \leq g_i(t) \leq \texttt{max\_dist}(t) - D_{1,i} && \forall i \in \{2, \dots, n+1\}\ \forall t, && (3.26)\\
&g_j(t) - g_i(t) \geq D_{i,j} + (x_{i,j}(t) - 1) * \texttt{max\_dist}(t) && \forall i, j \in \{2, \dots, n+1\},\ \forall t. && (3.27)
\end{aligned}$$

where $h_i(t)$ and $g_i(t)$ capture the incremental health facility count and travel distance added with each stop, respectively, and $\texttt{max\_stops}(t)$ and $\texttt{max\_dist}(t)$ denote the maximum number of locations that can be visited and the maximum distance it is feasible to travel as part of a single route, respectively.

Network Flow Constraints. The transition of samples through the eight stages of the system is described by equations (3.1)-(3.5). These transitions are linearized in the optimization model, using the following network flow constraints:

$$s_{i,k}(t) = s_{i,k}(t-1) - f_{i,k+1}(t) + f_{i,k}(t), \qquad \forall i, \forall t, \forall k \in \{1, \dots, 7\}, \tag{3.28}$$

$$\bar{M}(y_i(t) - 1) \leq f_{i,k}(t) - s_{i,k-1}(t-1) \leq \bar{M} y_i(t), \qquad \forall i, \forall t, \forall k \in \{2, 8\}, \tag{3.29}$$

$$\bar{M}(y_{n+1}(t) - 1) \leq f_{i,k}(t) - s_{i,k-1}(t-1) \leq \bar{M} y_{n+1}(t), \quad \forall i, \forall t, \forall k \in \{4, 6\}, \tag{3.30}$$

$$f_{i,k}(t) = \sum_{\tau \leq t} \alpha^k(\tau, t) f_i^{k-1}(\tau), \qquad \forall i, \forall t, \forall k \in \{3, 5, 7\}. \tag{3.31}$$

where $\alpha^k(\tau, t)$, in (3.31), represent the proportion of samples or results arriving at stage $k-1$ on day τ that progress to the next stage on day t. For example, if the district hub processes all samples on the next business day after arrival, then $\alpha^3(\tau, t) = 1$ if t is the next business day after τ and $\alpha^3(\tau, t) = 0$ otherwise.

Scheduling Constraints. Constraints on the auxiliary schedule variables $y_i(t)$ can be implemented to control when and how often each facility should be visited. For example, schedule constraints can be used to enforce minimum visit frequencies at a facility, specify a deadline for the next visit to a particular site, or prevent repeated visits to the same location on consecutive days. See Appendix B.1 for a full list of the additional scheduling constraints that were employed during implementation.

The OST Scheduling and Routing Model. Summarizing the above derivation, we obtain the following MIP;

OST Scheduling and Routing MIP :

$$\text{Minimize (3.16)} \tag{3.32}$$

Subject to:

Distance budget constraints: (3.17)–(3.19)

Routing constraints: (3.20)–(3.27)

Network flow constraints: (3.28)–(3.31)

Optional scheduling constraints: Appendix B.1

$f_{i,k}(t), s_{i,k}(t), h_i(t), g_i(t) \in \mathbb{R}_+;\ y_i(t), x_{i,j}(t) \in \{0,1\}.$

3.5.2 Menu-Based Routing Variants of the Deterministic Formulation

In practice, the most computationally expensive component of the model (3.32) is the construction of feasible routes according to constraints (3.17)–(3.27). The elimination of schedules that exceed the total weekly distance constraints (3.19) is particularly challenging, as it is not possible to determine whether a schedule violates these constraints until routes have been constructed for almost every day of the planning horizon. A disproportionate amount of time is spent constructing routes that meet the daily feasibility constraints, but are not feasible in combination with the remaining days in the planning horizon. We therefore introduce a more computationally efficient version of the model, in which some routes during the planning horizon are selected from a predetermined menu of feasible options.

Consider a menu of routes, indexed by $m \in \{1, \dots, M\}$, and let $\{r^1, \dots, r^M\}$ be binary vectors of length $n+1$, where $r_i^m = 1$ if location i is included in route m. Furthermore, let d^m be the distance of route m. Let $B^m(t)$ be a binary decision variable indicating whether route m is selected for day t. With this notation, we generate three additional constraints; one to ensure that the daily distance travelled is the sum of the individual route lengths, a second to ensure that the number of routes selected on day t does not exceed the number of available couriers, and a third to ensure that facilities are only visited if they are on a selected route;

$$d(t) = \sum_{j=1}^{M} d^j B^j(t), \tag{3.33}$$

$$\sum_{j=1}^{M} B^j(t) \leq \texttt{couriers}(t), \tag{3.34}$$

$$y_i(t) \leq \sum_{j=1}^{M} B^j(t) r_i^j(t), \qquad \forall i. \tag{3.35}$$

Each constraint (3.33)–(3.35) is then included in the network flow formulation for each day t for which we do not wish to generate routes as part of the optimization, but select them from the menu. This leads us to the following formulation, in which routes and schedules are fully flexible for days $t \in \{1, \ldots, \hat{t}\}$ but routes are selected from the menu for days $t \in \{\hat{t}, \ldots, T\}$;

OST Scheduling and Menu-Routing MIP :

Minimize (3.16) (3.36)

Subject to:

Flexible route budget constraints: (3.17) $\forall t \in \{1, \ldots, \hat{t}\}$, (3.18), (3.19)

Menu route distance budget constraints: (3.33) $\forall t \in \{\hat{t}, \ldots, T\}$

Flexible routing constraints: (3.20)–(3.27) $\forall t \in \{1, \ldots, \hat{t}\}$

Menu routing constraints: (3.34), (3.35) $\forall t \in \{\hat{t}, \ldots, T\}$

Network flow constraints: (3.28)–(3.31)

Optional scheduling constraints: Appendix B.1

$f_{i,k}(t), s_{i,k}(t), h_i(t), g_i(t) \in \mathbb{R}_+; B^m(t), y_i(t), x_{i,j}(t) \in \{0, 1\}.$

Table 3.1: Outline of the rolling horizon OST Solution Approach

The OST Solution Approach

DEFINE GLOBAL INPUTS:

1. Set length of flexible routing period: $\hat{t}$.
2. Set length of planning horizon: T.
3. Initiate *Global Route Menu* with existing fixed routes.

DAILY ROUTE GENERATION PROCESS:

1. Update input variables for health facility ST demand ($\mathbf{u}$):
- Set equal to USSD reported volume, if available.
- Set equal to expected volume, otherwise.
2. Update input variables for ST demand at other stages ($\mathbf{v}$):
- Retrieve data from external data sources (Appendix B.3.2).
3. Generate *Daily Route Menu*:
- Select the most common routes for the current weekday from *Global Route Menu*.
4. Solve **OST Scheduling and Menu-Routing MIP** (3.36):
- Key inputs: $\mathbf{u}$, $\mathbf{v}$, *Daily Route Menu*.
- Parameters: $\hat{t}$ and T.
5. Finalize next day routes (i.e., for $t = 1$):
- Save routes to *Global Route Menu*.
- Implement routes.

3.5.3 An Optimization-Based Approach to the Stochastic M-DMVRP

Our solution approach to the stochastic M-DMVRP problem relies on repeatedly solving the OST Scheduling and Menu-Routing MIP (3.36) over a rolling horizon. An outline of the procedure is provided in Table 3.1. In each period, we update the input variables to reflect the most up-to-date information about the location of samples in the network, which determines the ST demand at each stage. The rolling horizon approach allows each day's scheduling decisions to be made based on the most recent data available, while also ensuring that the model's solutions prioritize the continuous flow of samples and results through the ST system rather than simply maximizing the number of items transported on the current day in a Greedy fashion.

The trade-off between the solution quality and the computational efficiency of this solution approach is mostly driven by the length of the planning horizon (T) and the length of the flexible routing period ($\hat{t}$). By increasing the planning horizon, the model can better anticipate likely routing and scheduling decisions in the future and incorporate those for the period at hand. By increasing the flexible routing period, the model is better able to adapt to the incoming information on a daily basis, since it has a longer period for which it does not have to rely on routes from the menu. In the practical implementation of this solution approach (described in Section 3.7 and Section 3.8) we used a planning horizon of two weeks and a flexible routing period of one week.

3.6 Numerical Performance of Solution Approach

In this section, we present computational results describing the performance of our OST solution approach. First, in Section 3.6.1 we compare it with the optimal fixed policy and a lower bound on the optimal MDP policy, for small ST networks. Second, in Section 3.6.2 we compare it with various benchmarks which are solvable for ST networks of a realistic size, where size is characterized by the number of clinics collecting samples. Finally, in Section 3.6.3 we explore how the performance of dynamic scheduling and routing policies depends on the level of health worker compliance with the information sharing platform, as measured by the proportion of health facilities report ST demand on a daily basis.

3.6.1 Comparing OST and Optimal Policy Bounds for Smaller Networks

Design of Small Network Experiments. We now explore how the performance of the OST solution approach (with varying lengths of the rolling horizon) compares with a bound on the optimal MDP policy performance and an optimal fixed schedule policy. The MDP Lower Bound is generated by solving a truncated version of the original MDP using value iteration (Appendix B.2 provides a full derivation of this bound). The *Optimal Fixed Schedule* is generated by solving the MDP in Section 3.4 with the added constraint that the same schedule of visits must be followed every week. In order to generate the bounds on optimal performance, we are limited to examining small instances of the problem. Specifically, we consider a diagnostic network with four health facilities, one courier, and assume that at most 3 facilities can be visited on each route. The expected weekly sample volumes at each facility, μ_i, are uniformly distributed over the interval $[2, 7]$.

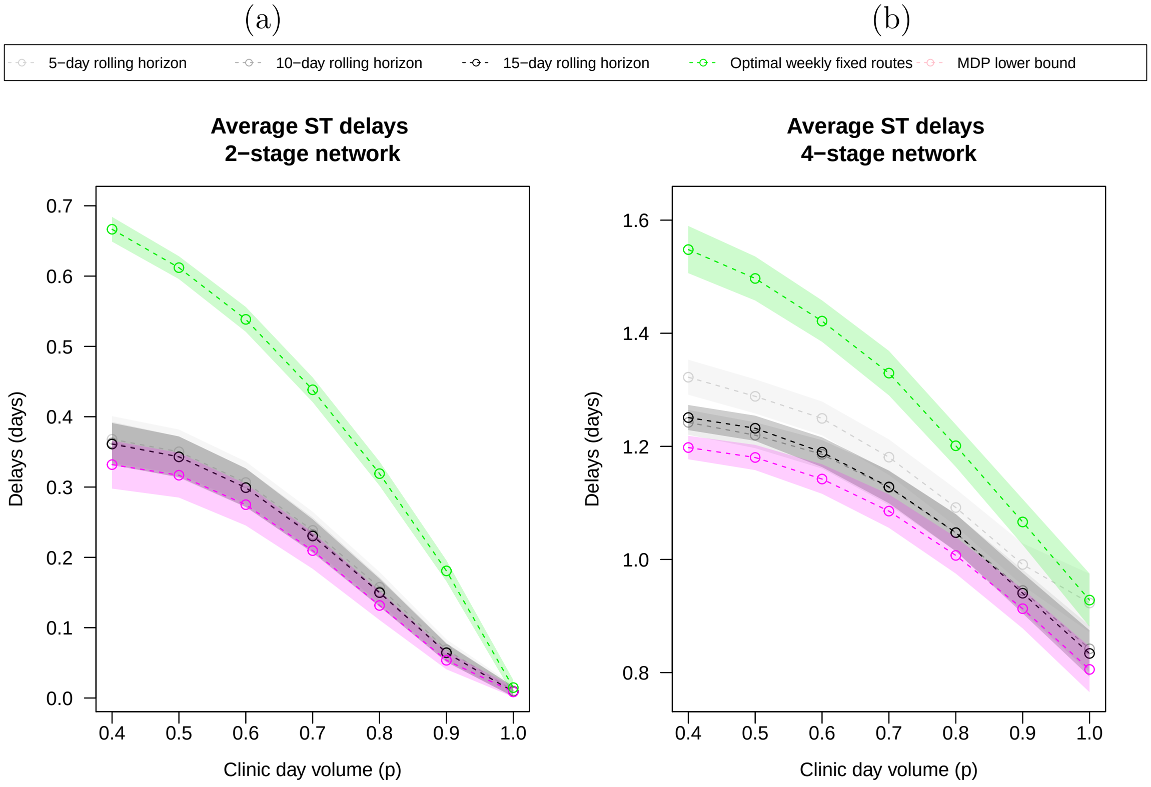
We report the sensitivity of our results along two dimensions. First, the ST sample arrival distributions ($u_i(t)$ in Section 3.4) are subject to predictable and unpredictable variability. The predictable variability (captured by the variable p) refers to the proportion of the total samples at each facility that are collected on each day of the week, in expectation (specifically on weekly "clinic days").[‡] Second, we present results for a 2-stage network (where samples only need to be transported from health facilities to district hubs) and for a 4-stage network (where samples must also be transported from the district hub to a molecular lab). For the four-stage problem, it is infeasible to transport samples to the molecular lab on the same day as visits are made to the health facilities.

Results. Figure 3.3 displays the results of this analysis (all points represent average performance over 20 simulations). Overall, we observe that as p increases, ST demand becomes more predictable and the average delays decrease for all methods, in both the 2-stage and 4-stage network. Focusing on the OST solution approach, we observe that a longer rolling horizon affects performance, but with diminishing returns. The performance of the 10- and 15-day rolling horizons is very similar, and both achieve on average 8% shorter delays than the 5-day rolling horizon. Intuitively, the rolling horizon OST solution approach significantly outperforms the optimal fixed schedule (by about 30%, on average), demonstrating the value of dynamic decision making in the system. More importantly, we observe from panel (b) that, on average, the 15-day rolling horizon application of the OST solution approach is within 7–20% of the lower bound on the optimal MDP policy in the 2-stage network and 1–5% in the 4-stage network. This demonstrates the strength of the OST solution approach, which is scalable and implementable, relative to the best achievable performance for the

[‡] We model the predictable variability using a parameter p, which denotes the proportion of samples that arrive on clinic days. The unpredictable variability refers to the stochastic noise around these conditional expectations, which we model using a Binomial distribution with parameters $(4, \mu_{i,w})$, where $\mu_{i,w} = \frac{p}{2}\mu_i$ on clinic days and $\mu_{i,w} = \frac{1-p}{3}\mu_i$ on non-clinic days.

system—particularly for multi-stage networks.

Figure 3.3: Performance of routing heuristics vs. proportion of samples on clinic days.



3.6.2 Comparing Scalable Solution Approaches for Larger Networks

Design of Large Network Experiments. We now compare the performance of the OST solution approach with various heuristics that are scalable and implementable for networks of the full eight stages and a realistic size. We generate synthetic diagnostic networks by placing the district hub a the center of a grid and placing $n \in \{10, 20, \ldots, 60\}$ health facilities randomly in the region $x \in [-1, 1], y \in [-1, 1]$. We assume a Euclidean distance matrix and fix a distance to the molecular lab of 2 units (a round trip of 4 units). We limit the maximum number of stops on a given to 5 and the maximum route distance to 3 (excluding trips to the lab). For each instance of the problem, we calibrate the weekly distance budget so that it is feasible to visit every health facility twice and the molecular lab once.[§] The arrival process is Poisson, with a higher rate on clinic days.[¶]

The objective of this analysis is to evaluate the performance of various heuristic solution approaches in a practical setting. We therefore fix the computational runtime which is available to each heuristic to 300 seconds per day. This reflects the practical consideration that since routes need to be evaluated and approved by ST coordinators in the field it is practically infeasible to implement a solution approach that requires excess runtime for every iteration. By constraining the computational resources available to each solution method we are able to highlight which solution methods are realistically scalable as the computationally intensive solution methods might perform well for small instances but potentially deteriorate for larger instances.

We compare the performance of the OST solution method to five benchmarks. First, a *Fixed Lab Schedule* heuristic which decomposes the lab visits from the health facility visits by fixing a weekly schedule for trips to the molecular lab and then solving the OST Scheduling and Routing MIP to generate routes and schedules for clinic visits on a daily basis.[‖] Second, a *Greedy Volume Heuristic* which optimizes the courier routes on a daily basis to transport as many samples as possible from one stage to the next, subject to penalties for samples left at facilities for longer than a week. Third, a *Fixed Weekly* heuristic which requires that the same set of routes and schedules are followed every week.[**] Fourth, a fully flexible *OST for Routing and Scheduling MIP* approach, which solves the full model (3.32) on a daily basis for the longest rolling horizon which is solvable in the computational runtime provided. Finally, an

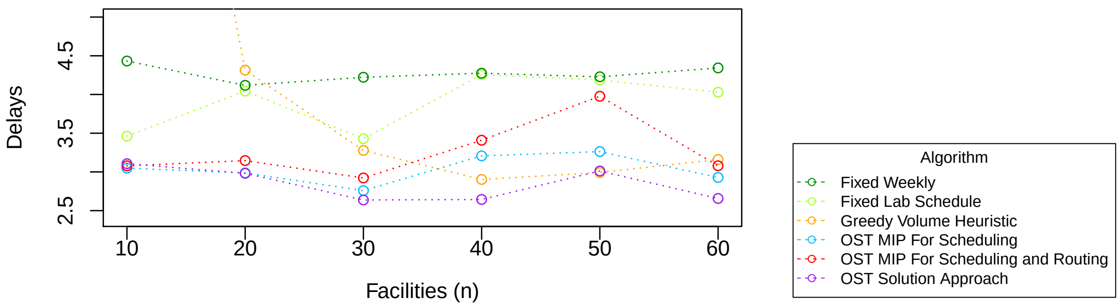
[§] Specifically, we solve a single period VRP problem in which all health facilities must be visited. We denote the total distance of all routes by d and the total number of routes by c. We then set the weekly distance budget for the multi-period problem to be $2d + 4$ (allowing for two health facility visits and a lab trip) and set the number of couriers to $\frac{c}{2}$ (rounding up to the nearest integer).

[¶] We simulate the arrival of samples at healthcare facilities using a Poisson distribution with mean $\lambda_{i,w} = \frac{\mu_i r_{i,w} p_i c_{i,w}}{\sum_{\tau=1}^{5} r_{i,\tau} c_{i,\tau}} + \frac{\mu_i r_{i,w}(1-p_i)(1-c_{i,w})}{\sum_{\tau=1}^{5} r_{i,\tau}(1-c_{i,\tau})}$, where μ_i are the expected weekly sample volumes at facility i, $c_{i,w} = 1$ if day w is a clinic day, and $r_{i,w}$ are random numbers from a uniform distribution. The μ_i values are drawn from an exponential distribution with mean 10, and each facility is randomly assigned two non-consecutive clinic days.

[‖] We explore all feasible fixed schedules to the molecular lab and display the best performing policy in each case.

[**] We generate optimal fixed routes for small instances and find high-quality solutions for large instances using a version of the OST Scheduling and Menu-Routing MIP for weekly routes.

Figure 3.4: Comparison of the performance of three routing heuristics in simulation experiments. Not shown: the greedy heuristic has average delays of approximately 13 days in small instances ($n = 10$).

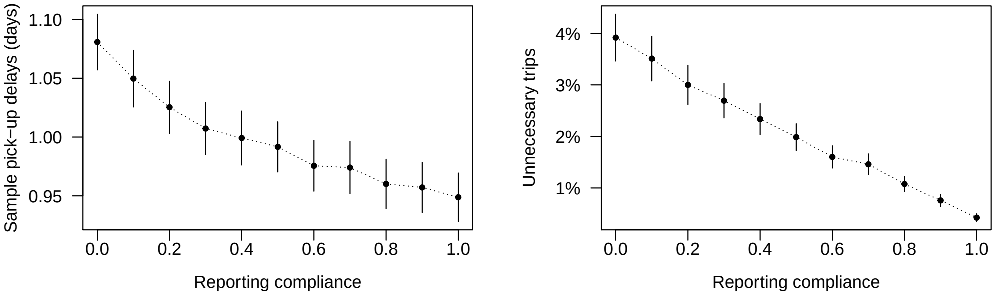


OST Scheduling MIP approach, which optimizes the visit schedule on a daily basis by solving the model (3.36) with $\hat{t} = 1$, thereby selecting routes from a menu for all periods.[††] For each algorithm, relevant parameters (such as rolling horizon length) were calibrated in preliminary simulations, and the final performance comparisons were based on 30 simulations of 120 days.

Results. Figure 3.4 presents the performance (average delays) of the 6 heuristics across different network sizes. First, we observe that, as expected, the *Greedy Heuristic* never has the best performance and the *Fixed Weekly* heuristic is always outperformed by the dynamic solution approaches. Second, comparing the *Fixed Lab Schedule* heuristic, which dynamically generates routes for only the health facility visits, to the multi-stage solution approaches, which dynamically generate schedules for health facility and molecular lab visits, demonstrates the value of explicitly modeling the multi-stage nature of the network. Third, among the three dynamic solution approaches, our OST Solution Approach has the best performance for all network sizes (except the smallest network, for which the *OST Scheduling MIP* has marginally lower delays, by 0.1 day on average). This demonstrates the value of being able to adjust the $\hat{t}$ parameter, which determines the duration of the fully flexible routing period, depending on the problem size. While the performance of each of the dynamic solution approaches is comparable for small instances, the computational benefits of the OST

[††] The initial menu of feasible routes is constructed in a training period.

Figure 3.5: Effect of reporting compliance rates on sample pick-up delays and unnecessary trips.



Solution Approach (while still allowing some flexibility in routing) result in improved performance for larger networks.

3.6.3 The Value of Information: Sensitivity to Information Sharing Compliance

Design of Numerical Experiments. As we describe in Section 3.2.2, the OST system comprises both an information sharing system and an optimization approach. The purpose of the information sharing system is to resolve short-term uncertainty about ST demand at each health facility. In this section we explore how the performance of the OST solution approach depends on the level of health worker compliance with the information sharing system. To this end, we simulate the performance of the OST solution approach on all the network instances described in Section 3.6.2, with the modification that only a randomly selected proportion p of the health facilities report their sample volumes through the information sharing system each day.

Results. Our results in Figure 3.5 indicate that the absence of sample volume reports leads to an increase of approximately 15% in the average sample pick-up delays (relative to 100% daily reporting), and that there are significant incremental improvements in sample pick-up times with increased reporting rates even if complete data is not available. The proportion of unnecessary trips included in the ST routes is also strongly correlated with the reporting compliance rates, with approximately 4% unnecessary trips when no sample volume data is available for healthcare facilities.

These results demonstrate that the availability of real-time sample volume data from health facilities improves the quality of optimized ST routes and schedules, and that incomplete data can still provide significant operational benefits.

3.7 Field Implementation

In this section we describe the field implementation of the OST system in Malawi. We start by discussing the objectives, design, and timeline of the implementation (Section 3.7.1) before describing the daily operational processes that were implemented (Section 3.7.2).

3.7.1 Overview of Implementation

Implementation Objectives. The primary objectives of the implementation were twofold. First, to evaluate the feasibility of the system in practice, both in terms of compliance with the USSD information sharing platform and in terms of operationally implementing the dynamic scheduling and routing of couriers on a daily basis. Second, to assess the practical impact of the OST system for reducing ST delays in a representative real-world setting.

Implementation Design. We selected three implementation districts—Rumphi, Salima, and Phalombe—that are representative of a range of R4H's typical ST operations in rural and semi-rural areas. Figure 3.6 and Table 3.2 describe the location and characteristics of the three districts. We observe that they are similar in terms of

Table 3.2: Pilot district summary statistics (2018)

District	Rumphi	Salima	Phalombe
Region	North	Central	South
Couriers	2	2	2
Facilities	18	18	15
KM per week*	1205	1148	894
Weekly samples	89	129	222
Population/km^2	50	222	325

*Average mileage based on fixed ST routes

Figure 3.6: Pilot district location

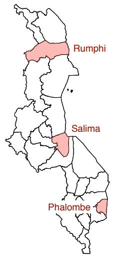
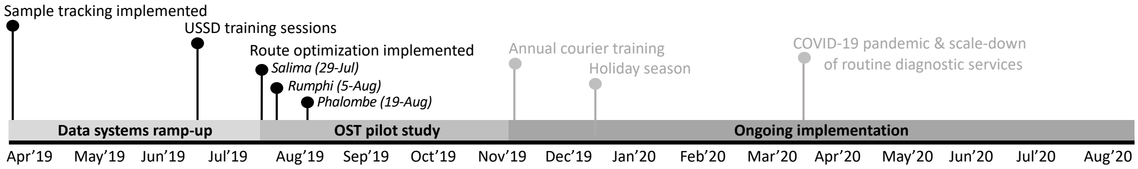


Figure 3.7: A timeline of the implementation of OST routes in Malawi



the number of health facilities (15 to 18) and are all serviced by two couriers. However, they also reflect the variation in demand for diagnostic services across different regions of the country. Population density and HIV prevalence are highest in the Southern region, resulting in high sample volumes and a higher density of health facilities. As a consequence, couriers in Southern districts generally travel shorter distances and transport higher volumes of samples and results. Towards the North, population density is lower and health facilities are geographically sparser, resulting in longer travel distances and lower sample volumes.

Implementation Timeline. The implementation timeline is depicted in Figure 3.7. We initially implemented the system for a *pilot phase* during July–October 2019.[‡‡] During this phase, significant emphasis was placed on minimizing potential disruptions to the diagnostic network. The OST routes were constrained to allow a gap of at most 7 days between successive visits to each healthcare facility, regardless of reported sample volumes, and all facilities received SMS notifications prior to scheduled visits.

Following the successful three month pilot phase, the system entered an *ongoing phase*, in which R4H elected to continue daily route optimization on an ongoing basis. During this phase of the implementation, greater flexibility was gradually introduced into the daily scheduling decisions in order to maximize the efficiency of the system. For example, healthcare facilities were no longer guaranteed visits on a weekly basis, which allowed the OST model greater scope to prioritize sites based on current sample volumes.

During 2020, operations within the diagnostic network were significantly disrupted by the COVID-19 pandemic. Routine diagnostic services such as viral load monitoring

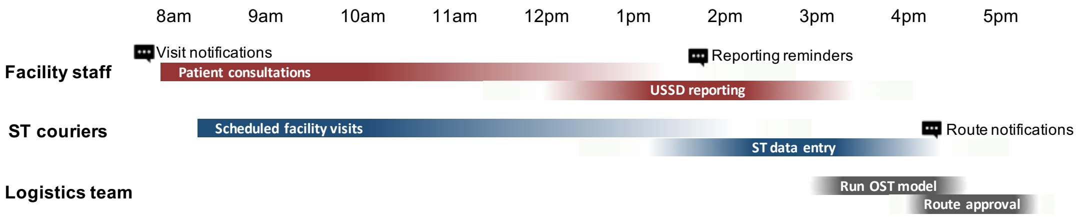
[‡‡] The implementation was conducted in collaboration with the Malawi Ministry of Health and was approved by Malawi's National Health Sciences Research Council.

were scaled back and resources within healthcare facilities and laboratories were redirected to support targeted testing and contact tracing. The OST system remained operational throughout these disruptions to continue sample transportation for essential diagnostic programs.

3.7.2 Daily Operational Process

The transition from fixed to OST routes was staggered over a three-week period starting in July/August 2019. We worked closely with R4H to develop a well-defined operational structure for managing daily route optimization, data collection, and internal communications, as illustrated in Figure 3.8. In summary, updated data was gathered on a daily basis (see detailed description of all input data sources in Appendix B.3). The route optimization was carried out by the research team between 3pm and 5pm each day to determine ST routes for the following day. To this end, model parameters and constraints were edited via a decision tool that allowed solutions to be tailored to reflect idiosyncratic daily developments in the field (e.g., courier absenteeism, fuel/travel restrictions, facility closures). Following route optimization the optimal routes were shared with regional sample transport coordinators for approval, to ensure that the routes were feasible, safe, and could be completed within a normal work day. The regional ST coordinators were also responsible for assigning each route to a specific courier and ensuring that the overall distribution of work was fair. A more detailed description of the personnel involved with the field implementation (along with a description of their responsibilities) is included in Appendix B.3.

Figure 3.8: The daily OST operational timeline



3.8 Field Impact

In this section we describe the impact of the OST system in Rumphi, Salima, and Phalombe. We first describe the impact on the primary outcome of ST delays (Section 3.8.1). We then describe the the effectiveness of the OST system in achieving two secondary objectives: improving data visibility (Section 3.8.2) and reducing unnecessary travel (Section 3.8.3).

3.8.1 Reduction of ST Delays

The primary objective of the OST system is to facilitate faster transportation of samples and results within the diagnostic network. To assess the performance of this approach, we report the average delays at each stage of the transportation network during the OST pilot period (August–October 2019). We assess the relative efficiency of the OST routes by comparing these results with the corresponding counterfactual delays associated with the fixed ST schedules in each district (see Appendix B.4).

The results in Table 3.3 are generated by analyzing the path of approximately 12,000 samples through the diagnostic network. The table demonstrates that the implementation of the OST system reduced average ST delays by approximately 25% (1.6 days) relative to the fixed ST schedules. The implementation of the OST system also altered the distribution of delays across different stages of the network. The most substantial reduction in ST delays was observed in the first stage of the sample transportation process: a 51% reduction in waiting times for picking up

Table 3.3: A comparison of average ST delays and standard deviation (in days) for fixed routes and OST routes.

		Average delays				Std. deviation	
Freq.*	Trip	Fixed	OST	Δ	%	Fixed	OST
0.75	Sample pickup (facility → hub)	2.26	1.10	-1.16	-51%	1.71	1.41
1.00	Sample delivery (hub → lab)	1.33	1.63	+0.3	+23%	1.46	1.43
0.62	Result pickup (lab → hub)	2.66	1.85	-0.81	-30%	2.02	1.33
0.77	Result delivery (hub → facility)	2.12	1.48	-0.64	-30%	1.46	1.37
	Frequency-weighted total:	**6.44**	**4.83**	**-1.61**	**-25%**		

*Proportion of observed samples/results that were transported between each set of locations.

new samples from healthcare facilities and a decrease of 0.3 days in the standard deviation of these delays. As a result of many samples reaching the district hub faster, the average amount of time spent waiting at the hub increased by 0.3 days. This increase is specifically linked to the timing of trips relative to the weekend (i.e., many samples that would otherwise have waited at healthcare facilities over the weekend were brought to district hubs on Friday). Average delays in the last two stages of the ST network (result pickup and result delivery) decreased by approximately 30%, and there was a substantial decrease (0.7 days) in the standard deviation of delays in fetching results from laboratories. We observe that the implementation of the OST system is associated with shorter average delays than fixed ST routes in all three implementation districts, with the most significant improvement in average delays in Salima (a reduction of 29%).

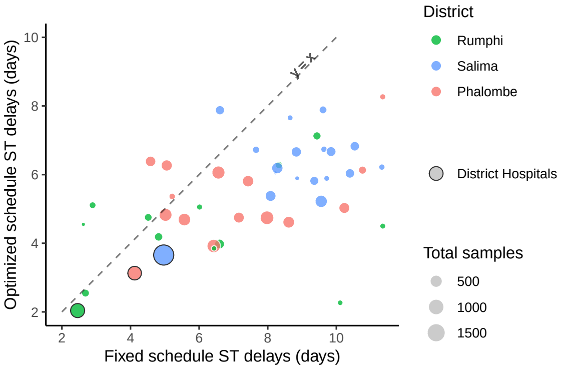
In Figure 3.9, we compare the average ST delays for healthcare facilities within the implementation districts. Implementing the OST system is associated with significantly shorter average delays at 31 of the 51 healthcare facilities ($p < 0.05$) and longer delays at 6 facilities, while delays at 14 facilities were not statistically affected. We observe that the range in delays across facilities was smaller following the implementation of the OST system (2–8.3 days) than for the fixed schedules (2.4–11.4 days), indicating that the OST system provides a more equitable distribution of resources across different locations.

Average travel distances during the pilot period were approximately 6% higher than the fixed schedules. Most of the excess travel occurred during the first month of optimized routing in each district. The increase in travel distances is partially attributed to inaccuracies in the initial distance matrices, which were continuously updated with data recorded by ST couriers. The sample couriers also had to adjust to unfamiliar routes and frequently modified their routes based on their knowledge of the local road conditions.

Table 3.4: Average ST delays by district

	Fixed	OST	% change
Rumphi	4.86	4.23	-13%
Salima	7.44	5.25	-29%
Phalombe	5.60	4.37	-22%

Figure 3.9: Average ST delays by facility



3.8.2 USSD Data Quality

Data quality is a key concern in the practical implementation of the OST system, as the model depends on up-to-date and reliable input data to determine where couriers should be dispatched on a daily basis. The initial pilot period (August – October 2019) placed significant emphasis on increasing the availability of this data by encouraging regular participation from all healthcare facilities. In subsequent months (November 2019 – August 2020) the field team shifted their focus to improving the accuracy of sample volume reports. Figure 3.10 illustrates the daily compliance rates with the USSD information sharing platform, which ranged from 53% during the initial ramp-up of the system in July–August 2019 to approximately 87% during the same period in 2020. Initial reporting rates were highest for viral load samples, and EID and TB reporting rates increased to similar levels by the end of 2019. Due to variations in services offered at different facilities, reporting "Other" sample volumes was optional during the initial pilot period.

Attaining this level of daily participation is particularly significant given that all reporting was voluntary and healthcare staff were not offered any financial incentives. It is also noteworthy that participation remained relatively high during March–June 2020, when many diagnostic services were disrupted by the COVID-19 pandemic.

Figure 3.10: USSD participation rates, July 2019 – August 2020

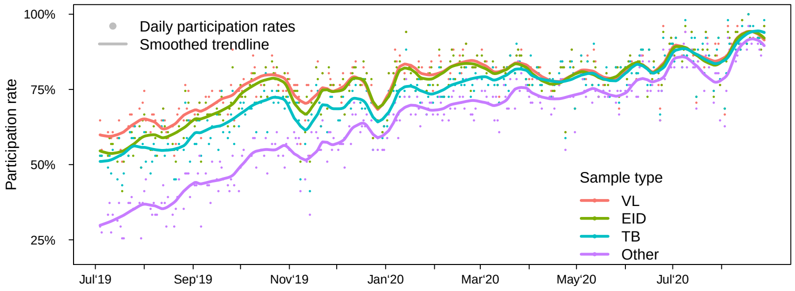


Table 3.5: Proportion of unnecessary visits in fixed vs. OST routes

	Fixed routes (Feb–Jun'19)	OST routes (Aug'19–Aug'20)
All districts	23.9%	10.6%
Rumphi	29.2%	19.9%
Salima	23.3%	4.7%
Phalombe	19.5%	8.6%

Appendix B.5 includes additional results on the accuracy of the reported data. By validating against the sample tracking records (collected by the couriers during health facility visits) we observe that approximately 85% of sample volume reports submitted via the USSD application were accurate, with the proportion of inaccurate reports decreasing over time, from 19% in July-August 2019 to 13% during the same period in 2020. Appendix B.5 also includes a discussion of the most frequent explanations for inaccurate reports.

3.8.3 Unnecessary Travel

Due to the improved data visibility within the ST network, the OST system significantly reduced the number of unnecessary trips that couriers made to healthcare facilities when there were neither samples nor results to be transported to or from that site. The average proportion of unnecessary trips across the three implementation districts was 24% for the fixed ST schedules, and decreased to 10.6% after the implementation of the OST routes (see Table 3.5). Appendix B.6 includes more detailed results about the reduction in unnecessary trips, including a summary of the reasons for unnecessary trips and a demonstration that unnecessary trips occurred more frequently at facilities with lower sample volumes, leading to the higher proportion of unnecessary visits in Rumphi.

3.9 Discussion

The practical implementation of OST routes generated several useful insights and provided ongoing opportunities to adapt and improve the system design. In this section, we briefly highlight three important factors that contributed to the successful implementation of the OST system and discuss their implications for the long-term management of ST operations.

Data Management. Managing daily data collection at healthcare facilities was the most practically challenging aspect of the OST system due to the large number of individuals involved in sample volume reporting. We worked closely with field staff during the pilot period to develop effective mechanisms for monitoring and improving data quality, and during 2020, R4H data clerks took over the day-to-day management of the USSD application. In order to ensure the long-term sustainability of the OST system, it is important that these data collection activities are integrated into R4H's ongoing ST operations and become an habitual activity at healthcare facilities.

Robustness. Although data quality improved significantly during the pilot period, missing and inaccurate data are an unavoidable occurrence given the scope and complexity of ST operations. It was therefore important to ensure that the daily logistics of the ST system were robust to communications failures and data errors. The results presented in Section 3.8.1 demonstrate that OST routes significantly improved the efficiency of ST operations in spite of the initial challenges associated with data accuracy and availability.

Flexibility. It was important for the OST model to be flexible enough to accommodate a variety of constraints that arose in day-to-day ST operations, including temporary logistical issues such as vehicle breakdowns, flooded roads, and staff absences. To accommodate these challenges, the OST model was incorporated into a decision tool that provided easy access to model constraints and parameters. We worked closely with the OST field manager and R4H regional coordinators to monitor the status of field operations in each district and incorporate input from couriers and field staff into the OST model.

Future Work and Expansion. From a practical perspective, R4H is currently developing plans to implement OST routes across the national ST program, based on the significant improvements in ST operations observed in Rumphi, Salima, and Phalombe. In preparation for this transition, R4H has already introduced the USSD application in approximately 300 additional healthcare facilities, and all R4H couriers have been trained to use the sample tracking application required to collect real-time input data for route optimization. From a research perspective, we believe the multi-stage version of the DMVRP introduced in this chapter motivates further research in the area of multi-stage transportation networks where the same vehicle capacity is shared between long-haul trips and last-mile delivery.

3.10 Conclusion

ST systems play an essential role in improving access to diagnostic services in many developing countries. In this chapter, we have addressed a common challenge observed in ST systems, namely, delays in the transportation of samples and results due to inefficient logistics. Our proposed solution consists of a comprehensive optimization model that determines daily ST routes in response the the demand for transportation at each location within the diagnostic network. In collaboration with R4H, we have implemented the optimized ST system in three districts in Malawi. The results of this implementation demonstrate that the OST system improves the efficiency of ST operations by reducing unnecessary travel and decreasing delays within the ST network.

During the last decade, many countries in sub-Saharan Africa have adopted hierarchical sample referral networks to facilitate large-scale VL monitoring programs in resource-limited settings. Our work provides a roadmap for increasing the responsiveness and flexibility of these systems in order to improve the efficiency of diagnostic networks.

3.11 Acknowledgements

This study was supported by Grand Challenges Award #OPP1182217 from the Bill and Melinda Gates Foundation.

The research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number U54EB027049. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Chapter 4

Optimal Deployment of Point-of-Care Instruments for HIV Viral Load Monitoring

4.1 Introduction

Since 2016, the World Health Organization has recommended viral load (VL) monitoring as the “the preferred monitoring approach to diagnose and confirm treatment failure” in HIV-positive patients taking Antiretroviral Therapies (ARTs) [232]. In resource-limited settings, blood samples collected from ART patients are generally transported to centralized laboratories for analysis, which can lead to delays of weeks or months in the availability of test results.

New technologies for Point-of-Care (POC) testing offer a faster alternative for VL monitoring that produces results within 1-2 hours, enabling clinicians to test and treat patients during the same visit. Early clinical studies of POC VL testing have demonstrated promising results, including better rates of retention and VL suppression [62], increased likelihood of switching to second-line therapies after a confirmed treatment failure, and shorter delays in initiating new treatments [149]. In light of these benefits, recent WHO guidelines [232] recommend the implementation of

POC testing in locations where it is feasible and cost-effective to do so, and highlight the need for further research to understand how to optimize the use of POC technology within the context of existing diagnostic networks.

In practice, POC testing can be considerably more expensive than centralized diagnostic testing due to the high cost of POC devices. The average cost of POC tests varies significantly based on the volume of samples collected at individual health facilities, and previous studies of VL monitoring programs in sub-Saharan Africa have found that many healthcare facilities do not collect enough VL monitoring samples to justify the substantial cost of POC testing equipment [81, 37].

Assessing the cost-effectiveness of POC testing within existing HIV treatment programs is challenging, as clinical studies usually take place in large, urban facilities [81, 74] which are not representative of the typical profile of healthcare facilities in resource-limited settings. Clinical studies generally focus on short-term individual outcomes (e.g., time to follow-up, earlier treatment switches, VL re-suppression) rather than broader benefits such as reduced HIV transmission, increased survival, and lower drug resistance in the general population. These studies also become rapidly outdated due to changes in HIV treatment policies—for example, it reasonable to expect that the benefits of POC testing may be somewhat smaller after the introduction of dolutegravir-based regimens and the scale-up of centralized testing in many sub-Saharan African countries, which have both lead to substantial increases in VL suppression rates.

In this work, we model and evaluate a range of policies for the implementation of POC VL monitoring in Malawi. Our policies address both capacity allocation (what types of devices should be used, and where should they be located?) as well as operational strategies for the use of POC testing within the context of the existing diagnostic network (how much testing should be performed at POC, and which patients should receive POC tests?). We consider differentiated approaches that are tailored to fit the demand for VL monitoring at each healthcare facility through a combination of centralized and POC testing, and we assess the incremental cost-effectiveness of these policies relative to the baseline policy of centralized testing alone. Our results demonstrate that while universal POC testing is unlikely to be feasible, optimized allocation and usage of POC instruments in combination with centralized testing is significantly more cost-effective.

4.2 Methodology

Our methodology consists of three key steps. First, we develop detailed cost estimates for POC VL monitoring at healthcare facilities in Malawi (Section 4.2.1). Second, we used an established simulation model to estimate the effectiveness of introducing varying amounts of POC testing in combination with centralized VL monitoring (Section 4.2.2). Finally, we used the data generated in the first two steps to formulate a mixed-integer optimization model that determines the optimal allocation of POC capacity to healthcare facilities in Malawi under a variety of different operational constraints (Section 4.2.3).

4.2.1 POC Cost-Per-Test Analysis

This section describes the methods used to estimate the cost-per-test of implementing POC VL monitoring at healthcare facilities in Malawi. Our analysis considers each healthcare facility individually, and provides facility-specific cost estimates that depend on both the demand for VL monitoring at the facility, as well as the cost, capacity, and utilization of POC testing equipment.

4.2.1.1 Input Data

Point-of-Care Instruments. Our analysis included four POC devices which are commonly used for VL testing in resource-limited settings. Three of these devices are GeneXpert systems manufactured by Cepheid [44], which can be configured with a varying number of modules to process multiple tests simultaneously (we considered configurations with 1, 2, 4, 8, 12, or 16 modules). We also included the Abbott m-PIMA Analyser [2], which processes a single sample at a time. We estimated the instruments' daily test capacity assuming approximately 7 hours of use per day (35 hours per week).

Table 4.1: POC instrument configurations and costs.

Instrument	Modules	Daily capacity	Costs	
			Fixed	Variable
GeneXpert II	1	4	$30,920	$19.25
GeneXpert II	2	8	$34,030	$19.25
GeneXpert IV	4	16	$42,840	$19.25
GeneXpert XVI	8	32	$83,504	$19.25
GeneXpert XVI	12	48	$102,944	$19.25
GeneXpert XVI	16	64	$120,354	$19.25
m-PIMA	1	6	$40,420	$29.10

We used information from manufacturer websites and recent literature [37, 151, 14, 142, 48, 203] to estimate both the fixed and variable costs associated with each instrument. Fixed costs included the cost of the device, accessories, installation, and maintenance over an expected lifespan of 5 years, and variable costs per test included test cartridges, electricity and utilities, sample collection kits and PPE, and staff labor. A summary of the instruments included in our analysis is provided in Table 4.1, and additional details on the data sources and cost assumptions are provided in Appendix C.1.1.

Healthcare Facilities in Malawi. We compiled a list of healthcare facilities offering VL testing in Malawi based on data extracted from the national Laboratory Information Management System (LIMS). This database contains records of all VL samples analyzed at 10 national molecular laboratories, which account for over 99% of VL tests conducted in Malawi [17]. During 2021, 702 healthcare facilities in Malawi referred at least one VL sample for testing at a molecular laboratory. The distribution of annual sample volumes across healthcare facilities was highly skewed (see Figure 4.1), with 75% of facilities recording fewer than 1000 samples during the year, and just 3 facilities recording over 10,000 samples. The top 10 facilities accounted for over 16% of national volumes.

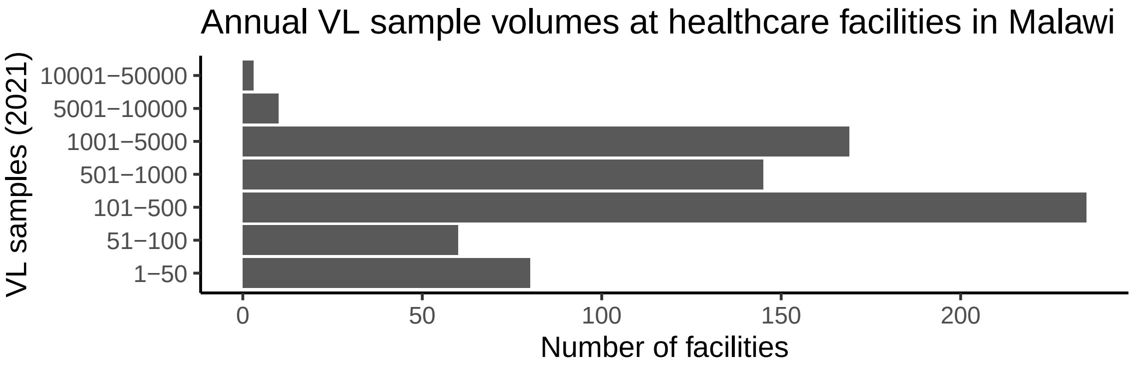


Figure 4.1: A summary of 2021 VL sample volumes at 702 healthcare facilities in Malawi.

4.2.1.2 Cost Model Calculations

Coverage and Utilization. We estimated the daily demand for VL monitoring at each healthcare facility based on the VL sample collection dates recorded in the national laboratory database during 2021. We then compared these daily volumes to the testing capacity of each POC instrument in Table 4.1 in order to determine how many VL samples would have been tested on the device each day. We assumed that all samples would be tested on the POC instrument on days when the demand for testing was lower than the device capacity, and that any excess samples would be referred for centralized testing on days when the total number of samples exceeded the device capacity. Using this data, we calculated the *POC coverage* rate (i.e., what proportion of the annual VL samples collected at the facility could have been tested on the device) and the utilization rate (i.e., the total number of POC samples processed divided by the total instrument capacity over a 1-year period).

Cost-Per-Test. For each facility, we calculated the cost per POC test associated with each of the POC instruments. The cost-per-test includes the variable costs per sample (Table 4.1), as well as the device fixed costs divided by the total number of POC samples tested on the instrument. For comparison, we also estimated costs for centralized testing at each facility, which ranged from $20.28–$22.28 depending on the transportation requirements. Full details of the centralized cost estimates are available in Appendix C.1.1.5.

4.2.1.3 Operational Strategies

We performed additional cost analysis to evaluate two operational strategies that have the potential to decrease POC costs and/or increase POC instrument utilization.

Expanded ART Clinic Schedules. Many healthcare facilities in Malawi only offer ART clinic services on specific days of the week, which leads to a high degree of variability in daily VL sample volumes (for example, [78] found that most healthcare facilities in three districts in Malawi collected more than 80% of their total sample volumes on just two days of the week). This uneven demand for VL monitoring is problematic in the context of POC testing, because POC instruments are likely to be idle on days when ART services are not offered, and the relatively high volume of samples during ART clinic days may require larger, more expensive instruments. To address these problems, we considered modifying facility schedules to provide ART services and VL monitoring 5 days per week.

To evaluate the potential impact of expanded ART clinic schedules at each facility, we simulated the adjusted daily demand for VL testing using a Poisson distribution with mean equal to 20% of the facility's weekly sample volumes. We then repeated the coverage and utilization analysis described above using the modified daily volumes.

Cost Sharing. The capacity of POC instruments for VL monitoring is significantly larger than the demand for VL monitoring at many healthcare facilities in Malawi, which leads to low utilization and high average costs. To address this issue, we considered sharing POC devices with other types of testing conducted in healthcare facilities, such as TB testing or HIV EID. In these scenarios, we assumed that a percentage of the instrument's fixed cost would be covered by other disease programs, and a corresponding percentage of the instrument's capacity would be used for other types of samples. We considered three levels of cost-sharing accounting for 25%, 50%, or 75% of the total instrument capacity (where 0% is the baseline scenario with no cost-sharing). We repeated the coverage, utilization, and cost analysis for each device using the reduced fixed costs and capacity corresponding to each level of cost-sharing.

4.2.1.4 Summary of POC Cost Model

Our final cost model estimated the cost-per-test, coverage, and utilization of each POC instrument at each healthcare facility in Malawi for both the baseline and expanded facility schedules, as well as varying levels of cost-sharing with other types of tests. Using this model, we were able to compare the performance of different instruments at each facility and select the most appropriate instruments under a variety of operational constraints (for example, minimum POC coverage targets). In general, we assumed that the "best" instrument for a particular facility was the option that met the operational requirements at the lowest cost-per-test.

4.2.2 POC Impact Analysis

We used the HIV Synthesis model to simulate the impact of implementing POC VL monitoring in combination with centralized testing at healthcare facilities in Malawi. The HIV Synthesis model simulates HIV transmission, progression, and treatment among a population of individuals over 3-month increments, and has previously been used to evaluate other HIV treatment policies such as the scale-up of VL monitoring [169] and the introduction of new ART drugs [175, 171]. Additional information and access to the model is available through the HIV Synthesis project website [168].

In our analysis, we simulated the impact of implementing different POC testing strategies at healthcare facilities in 2023, for a total period of 5 years. This time frame was selected to reflect the average lifespan of POC instruments. The key strategies that we evaluated were (1) what proportion of tests should be analysed on POC instruments (vs. at centralized laboratories), (2) which patients should receive POC tests, and (3) whether POC tests would be processed at the facility, or near-POC (i.e., on a POC instrument located at a nearby facility). The sections below describe the assumptions made about each of these strategies, and how they were implemented in the HIV Synthesis model. Additional details of the model calibration for Malawi's population are described in Appendix C.2.2.

4.2.2.1 ART Treatment Policies

Based on Malawi's HIV treatment guidelines [128, 129] we simulated dolutegravir-based ART regimens as the preferred treatment for all adults diagnosed with HIV, with protease inhibitor regimens available to patients who switch drugs due to treatments failures or adverse reactions. We assumed that patients on ART should receive routine VL testing at 6 and 12 months after starting treatment, and every 12 months thereafter. Patients with a high VL result during routine VL monitoring received treatment adherence counseling, followed by a confirmatory VL test 3-6 months later. Treatment switches to second- or third-line therapies were considered for patients with a high VL result in the confirmatory tests (i.e., after two consecutive high VL results in a 12-month period).

4.2.2.2 VL Testing Assumptions

Each VL monitoring test in the simulation was modeled as either a centralized test, a POC test, or a near-POC test. Prior to 2023, all VL monitoring was done via centralized laboratories. During the intervention period (2023–2028), we considered a variety of scenarios in which the proportion of samples analysed on POC instruments ranged from 5% to 90% (with the remaining samples modeled as centralized tests). We also considered three "all-or-nothing" scenarios: 100% centralized testing (the baseline scenario), 100% POC testing, and 100% near-POC testing. In scenarios which used a combination of different types of tests, we only considered two options: [centralized + POC testing] or [centralized + near-POC testing]. We did not consider a combination of POC and near-POC testing, as it would be inefficient for facilities capable of performing POC testing to send samples to other locations for near-POC testing.

All types of VL monitoring tests were modeled with similar sensitivity and specificity, and the same treatment guidelines and drug regimens were followed regardless of the type of test administered. However, POC tests were assumed to have higher success rates and faster follow-up than centralized tests, and patients were more likely

Table 4.2: Summary of key parameters and assumptions for POC, near-POC, and centralized testing. These ranges represent steady-state values after 2025.

	Centralized	POC	Near-POC	
Successful result probability	87%	96%	91.5%	
			$p = 0.8$	$p = 0.2$
Follow-up time	3 months	$<$ 7 days	<7 days	3 months
Respond to adherence counseling[1]	$1 - \rho$	$1 - \rho/2$	$1 - \rho/2$	$1 - \rho$

[1] Patients who respond to adherence counseling increased their baseline treatment adherence for 6 months. The parameter ρ took values (0.65, 0.3, 0.1) with probability (0.15,0.7,0.15), based on the original HIV Synthesis model.

to respond to adherence counseling provided soon after testing. Near-POC tests were assumed to offer similar advantages to POC tests, but to a lesser extent. A summary of these assumptions is provided in Table 4.2.

The assumptions used to model POC and near-POC testing are based on data from a variety of empirical sources, including POC pilot studies in Chiradzulu [149] and Lilongwe [77] districts in Malawi, the STREAM trial [62], and aggregated reports on POC VL monitoring programs in various locations in sub-Saharan Africa [29]. When necessary, we extrapolated trends from observed data to account for the scale-up and improvement of diagnostic services over time [117]. A detailed discussion of the parameters used to model VL testing is provided in Appendix C.2.3.

4.2.2.3 Priority Strategies

In scenarios using a combination of centralized and POC VL monitoring, we considered four different strategies for deciding which patients should be prioritized for POC testing. The baseline strategy (**R**) was to allocate tests randomly among all patients receiving VL monitoring, and the second strategy (**WHO**) prioritized key groups of patients identified by the 2021 WHO guidelines, including individuals who were pregnant; under the age of 20; returning for a follow-up test after a high VL result; diagnosed with TB, opportunistic infections or other HIV-related complications; or undergoing the first VL test after re-entering care [233]. We also compared two additional strategies that prioritized similar high-risk groups, but with greater emphasis on specific interventions and outcomes. The third strategy (**H/F**) gave greater priority to individuals with a recent history of high VL or suspected treatment failure, while the fourth strategy (**T/A**) assigned higher priority to individuals at higher risk of transmission or poor treatment adherence (specifically, individuals who are pregnant, recently started ART, have more than one (unprotected) sexual partner, or are under the age of 30). A more detailed description of each strategy is available in Appendix C.2.3.6.

4.2.2.4 Simulation Output

The main outcome of interest in the simulation model was the total Disability Adjusted Life Years (DALYs) for individuals aged 15-80 years old. For individuals living with HIV, DALYs were impacted by opportunistic infections, drug toxicity, and death from HIV-related causes. We also used the simulation output to estimate the total costs associated with HIV-related healthcare services, including treatment costs for people living with HIV (clinic visits, ART drugs, adherence counseling, etc.), as well as broader health system costs such as HIV testing, PrEP, and TB monitoring and treatment. A more detailed description of the cost and DALY calculations is available in [169] and the associated supplementary material.

4.2.2.5 Facility Impact Estimates

Our impact analysis was based on 550 unique simulation instances which were each used to evaluate 90 different testing scenarios, including varying proportions of either POC or near-POC testing allocated according to the four priority strategies described above. To estimate the impact of these strategies at different healthcare facilities in Malawi, we matched each facility to at least 20 simulation instances with similar HIV prevalence rates in the initial population. We aggregated the costs, DALYs, and other outputs across each of the selected simulations and then scaled the aggregated output to an appropriate size for the population of each healthcare facility. Details of the facility HIV prevalence estimates are provided in Appendix C.1.2.1 and the approach used to scale the simulation output is described in Appendix C.2.2.1.

4.2.3 Optimization Model

The facility cost and DALY estimates described in Sections 4.2.1–4.2.2 were used to formulate a mixed-integer optimization model to determine the most effective national strategy for implementing POC VL monitoring in Malawi. In this section we give a general overview of the model assumptions and constraints without any mathematical notation. The full model formulation is provided is Appendix C.3.

4.2.3.1 National Policy Choices

The optimization model includes several high-level operational decisions which apply to all healthcare facilities. National policy options include whether to allow near-POC testing, which patients to prioritize for POC testing, whether to expand ART clinic schedules and/or allow cost-sharing of POC instruments, and what proportion of national sample volumes should be analysed on POC devices. A summary of these options is provided in Table 4.3. The national policy decisions can either be provided as inputs to the optimization model to find solutions that meet specific operational constraints, or left as decision variables to be optimized.

4.2.3.2 Facility Testing Strategies

For each healthcare facility, the optimization model determines what proportion of the facility's samples should be tested on POC instruments, what type of instrument should be used, and whether the facility will refer samples for near-POC testing or test near-POC samples from other facilities. A summary of these decisions is provided in Table 4.4. Note that all facility-level strategies must comply with the national policy decisions described above. The optimization model also includes a number of general feasibility constraints to ensure that the testing strategy implemented at each facility is compatible with the type of instrument allocated (i.e., there must be sufficient POC or near-POC capacity to test the corresponding proportion of samples).

Table 4.3: High-level operational strategies considered in the optimization model.

Strategy	Options
Type of tests	Centralized + POC Centralized + POC or Near-POC
Facility schedules	Baseline (limited ART clinic hours) Expanded (testing 5 days per week)
Priority strategy	**AoN** (All-or-Nothing) — each facility offers one type of testing **R** — POC tests assigned randomly **WHO** — POC tests assigned according to WHO priorities **H/F** — high priority for previous high VL / suspected treatment failure **T/A** — high priority for transmission / adherence risks
Cost-sharing	0% — no cost-sharing 25%, 50%, 75% — POC instruments shared witho other types of testing
POC coverage	0-100% — proportion of national samples tested on POC instruments

Table 4.4: Facility-level decisions in the optimization model.

Strategy	Options
Type of tests	Centralized only POC + centralized Near-POC + centralized
(near-)POC coverage	5%, 10%, 20%, ..., 90%, 100% — the proportion of samples at the facility that will be tested on POC instruments
Instrument & POC capacity	Instruments in Table C.1, adjusted to reflect capacity shared with other programs
Near-POC hub	Will this facility test samples from other locations?

4.2.3.3 Objective

The objective of the optimization model is to maximize the total DALYs averted relative to the baseline scenario of 100% centralized testing at all healthcare facilities. The total DALYs averted is calculated by aggregating the sum of the estimated DALYs averted at each healthcare facility, based on the VL monitoring strategy selected for the facility. A key assumption in these calculations is that the DALYs averted at each facility are independent of the strategies selected at other sites.

4.2.3.4 Costs and Cost-Effectiveness

In the absence of any cost constraints, the optimization model will always recommend implementing as much POC testing as possible in order to obtain the largest impact. It is therefore necessary to constrain either the costs or the cost-effectiveness of the model solutions. We have selected the latter approach, which is guaranteed to return solutions with an incremental cost-effectiveness ration (ICER) less than or equal to $500 per DALY averted (if any feasible solutions exist). This threshold is frequently used as a standard for cost-effective interventions in the context of HIV treatment programs in sub-Saharan Africa [174, 172, 169, 229].

Cost Calculations. The total costs calculated in the optimization model include four separate components: (1) POC instrument costs, which depend on the type of device allocated to each facility and are based on the estimated 5-year fixed costs in Table 4.1; (2) (near-)POC variable costs, which are calculated by multiplying the estimated POC sample volumes over a 5-year period (based on simulation output) with the cost-per-test estimates in Table 4.1, as well as transportation costs for near-POC samples (see Appendix C.1.1.6); (3) centralized testing costs, which are calculated by multiplying the estimated centralized testing volumes over a 5-year period by the estimated cost per centralized test ($20.28 – $22.28, see additional details in Appendix C.1.1.5); and other health system costs, based on the simulation estimates associated with the testing strategy selected for each healthcare facility.

Cost-Effectiveness Constraints. We require that all POC devices allocations in the optimization model must have an ICER of at most \$500 per DALY averted. We use the baseline scenario of 100% centralized testing as a reference point for these calculations, and we calculate the ICER by comparing the incremental difference in estimated costs and DALYs associated with the testing strategy selected for each facility. In cases where near-POC testing is used, we require the total ICER for all facilities sharing the same POC instrument to be less than or equal to \$500. At facilities where near-POC testing is not used, we apply the ICER constraint to each facility individually.

Applying the cost-effectiveness constraints to each POC instrument is a more stringent requirement than simply requiring that the total national costs and DALYs have an acceptable ICER. The motivation for using these more stringent constraints is to ensure that the national allocation strategy is fair and consistent. From a practical standpoint, this means that we cannot use the benefits gained from a POC instrument in one facility to justify the cost of placing an instrument in another facility where it will not be used cost-effectively.

4.2.3.5 Realistic Policy Assumptions

The combinatorial structure of the optimization model can be used to model a wide range of national policies (see Table 4.3) and find the corresponding optimal allocation of POC instruments associated with each scenario. In our analysis, we focused primarily on a narrower range of policies which are likely to be compatible with Malawi's existing diagnostics infrastructure. In particular, we performed a detailed analysis of optimal allocation strategies to transition approximately 20% of VL testing to POC instruments, with the remaining 80% of tests processed at centralized laboratories. This combination of POC and centralized testing represents a realistic short-term target that has the potential to generate significant impact, while also maintaining relatively high utilization of existing centralized infrastructure. Related work by [80] used a similar target of 15% POC coverage in South Africa.

Near-POC Network Structure. We also imposed several constraints on near-POC testing strategies in order to take advantage of existing infrastructure and logistics within the current diagnostic network. We assumed that all near-POC testing would be conducted at district hospitals, as these facilities are likely to have the appropriate infrastructure and staff to provide near-POC services (many histrict hospitals already perform testing for TB and HIV EID). District hospitals also serve as hubs in Malawi's sample transportation network, which would allow transportation of POC samples via similar channels to centralized samples, and at similar costs.

If near-POC testing is implemented in a district, we required that each facility in the district must have equal access to near-POC testing capacity (i.e., all facilities should send a similar proportion of their samples for near-POC testing) unless the facility is allocated its own POC device. We assumed that there is a small fixed cost of $200 per year for each facility that implements near-POC testing, and that transportation costs for near-POC samples are similar to those of centralized samples. In line with the assumptions in Section 4.2.2.2, we assumed that facilities allocated an instrument for POC testing could not refer additional samples for near-POC testing at other facilities (i.e., any samples that exceeded the capacity of the allocated instrument were referred for centralized testing).

4.3 Results

In this section we provide three sets of results that correspond to the cost, impact, and optimization analysis described in Section 4.2.

4.3.1 Cost of POC Testing at Healthcare Facilities in Malawi

We used the cost model described in Section 4.2.1 to estimate the cost-per-test for POC VL monitoring at each healthcare facility in Malawi in a range of scenarios, including 100% POC testing vs. combined POC and centralized testing, baseline clinic schedules vs. expanded clinic schedules, and varying levels of cost-sharing with other types of samples. In each scenario, we estimated the cost-per-test at each facility by selecting the POC instrument that fulfilled the operational constraints at the lowest cost.

In general, the distribution of testing costs across healthcare facilities was highly skewed (see Figure 4.2), with a small number of high-volume facilities (primarily district and central hospitals) able to perform POC testing at comparable costs to centralized testing, while the vast majority of facilities had significantly higher costs.

Table 4.5: The average cost-per-test, including fixed and variable costs, for POC VL monitoring at healthcare facilities in Malawi.

	Shared	Baseline schedules			Expanded schedule		
	capacity	Cost	Coverage	Utilization	Cost	Coverage	Utilization
POC only	0%	$32.84	100%	16%	$26.56	100%	42%
	25%	$31.26	100%	16%	$25.24	100%	43%
	50%	$30.00	100%	16%	$23.84	100%	45%
	75%	$28.20	100%	17%	$22.64	100%	45%
POC + centralized	0%	$28.53	78%	31%	$26.27	90%	58%
	25%	$26.93	72%	34%	$24.62	88%	64%
	50%	$25.35	62%	37%	$23.03	82%	72%
	75%	$23.18	51%	47%	$21.42	74%	80%

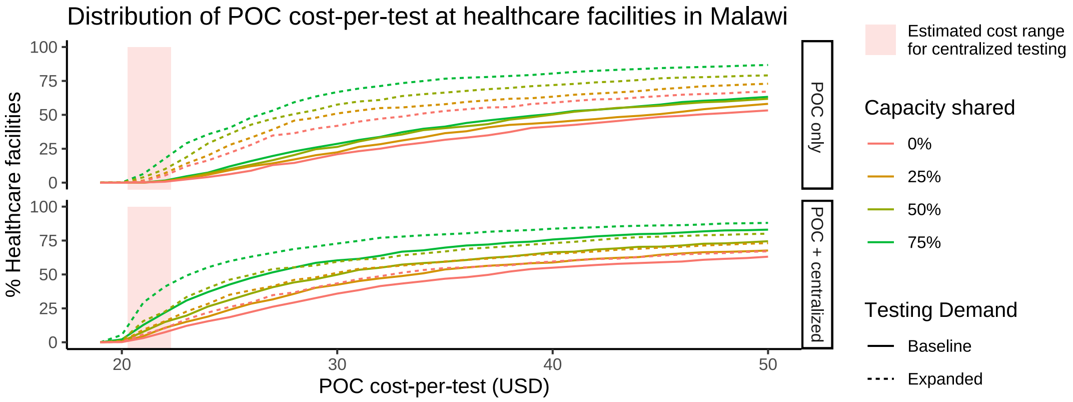


Figure 4.2: The percentage of healthcare facilities able to perform POC VL monitoring at varying cost-per-test thresholds. The solid red line represents a baseline scenario in which all samples are collected on designated clinic days according to 2021 schedules, and the POC device is used only for VL monitoring.

100% POC Testing. The top four rows in Table 4.5 and the first graph in Figure 4.2 show the estimated costs for 100% POC testing (i.e., each facility must have a POC instrument with enough capacity to process all VL samples). When samples were collected according to the baseline clinic schedules, 100% POC testing was more expensive than centralized testing at every healthcare facility and the national average cost-per-test was $31.88, with an average of only 16% utilization of POC capacity. The estimated costs for 100% POC testing were lower for the expanded clinic schedules (i.e., when samples were collected every day of the week), with a national average of $26.27 per test and 42% capacity utilization. For both the baseline and expanded schedules, sharing POC instruments with other types of testing reduced the average POC costs, but did not have a significant impact on utilization.

POC + Centralized Testing. The last four rows in Table 4.5 and the second graph in Figure 4.7 show the corresponding cost estimates for implementing POC testing in combination with centralized testing at every facility (i.e., any excess samples that exceed the POC instrument's capacity are referred to a centralized laboratory). In this case, the average cost per POC test was $28.53 (31% utilization) using the baseline schedules, and $26.27 (58% utilization) if samples were collected every day of the week. Sharing POC instruments with other types of testing reduced costs and increased utilization, but resulted in a greater number of samples referred to centralized laboratories.

POC Instruments. The type of POC instruments selected in each scenario is summarized in Figure 4.3 and Table C.3. Note that the m-PIMA instrument was not selected in any of the scenarios considered, primarily due to the higher cost of test cartridges. Scenarios using 100% POC testing required significantly larger instruments in order to meet the peak demand for VL testing at each facility, as did scenarios where samples were collected according to the baseline clinic schedules, and scenarios where testing capacity was shared with other types of samples. In the scenarios where testing was performed every day without any capacity sharing, the maximum capacity GeneXpert configuration (16 modules, 64 tests per day) was selected for a single facility (Bwaila Hospital, $> 30,000$ samples per year), while the 12-module configuration was selected for three other large facilities with 9,000–13,000 samples

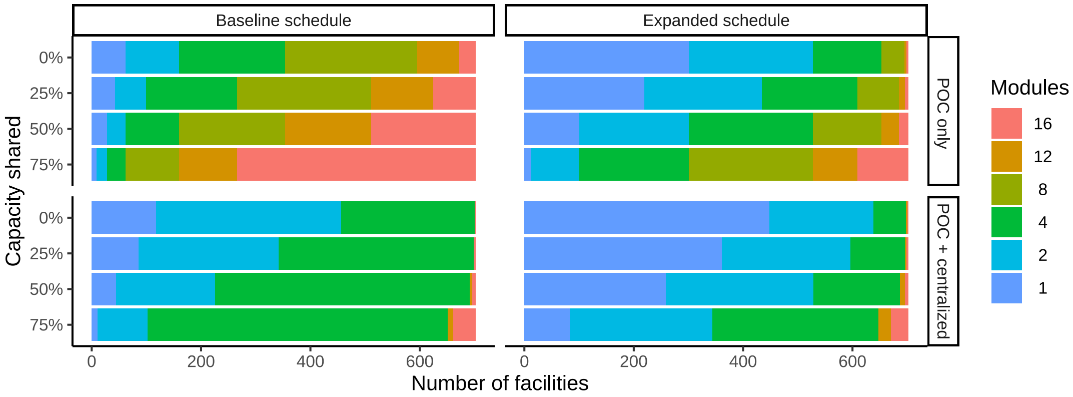


Figure 4.3: A summary of the number of POC test modules allocated to healthcare facilities for the operational strategies in Table 4.5. Each test module is assumed to provide 4 tests per day. The GeneXpert instruments corresponding to each number of modules are listed in Table 4.1.

per year.

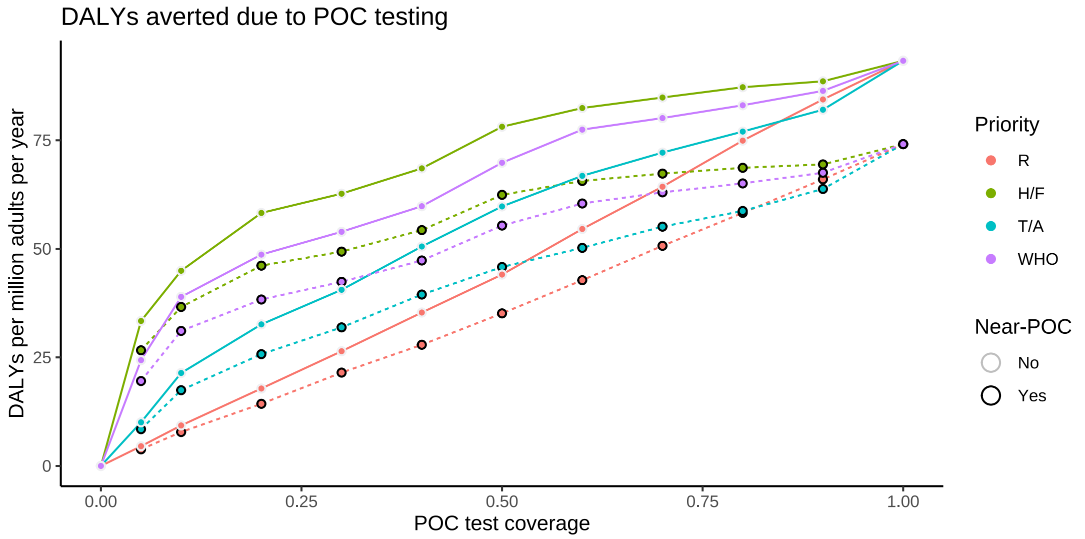
4.3.2 Impact of POC Testing

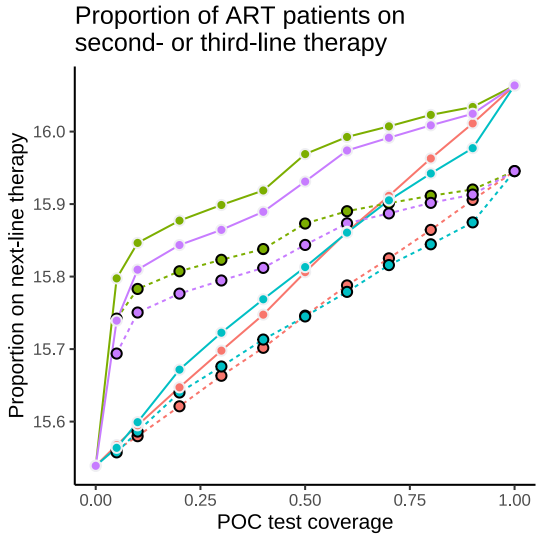
Using the HIV Synthesis model, we simulated various strategies for introducing POC and near-POC testing at healthcare facilities in Malawi. The results presented in this section represent the aggregated outcomes across all healthcare facilities and are scaled to match the approximate size and composition of Malawi's population, as described in Appendix C.2.2.

DALYs Averted. Figure 4.4 shows the simulated impact of testing varying proportions of VL samples on POC instruments using the priority strategies introduced in Section 4.2.2.3. Impact is measured in terms of DALYs averted relative to the baseline strategy of 100% centralized testing. Implementing 100% POC testing for VL monitoring resulted in the largest impact, approximately 95 DALYs averted per million people per year, while 100% near-POC testing averted approximately 75 DALYs per million people per year.

For combined POC and centralized testing, the estimated impact varied based on the priority strategy used to determine which patients were allocated POC tests. For random allocation of POC tests (**R**), the number of DALYs averted increased linearly

Figure 4.4: Simulation results: DALYs averted due to POC testing on over a five-year period.





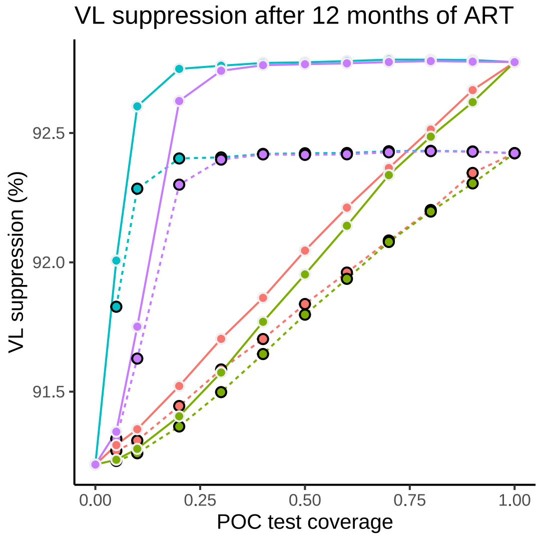
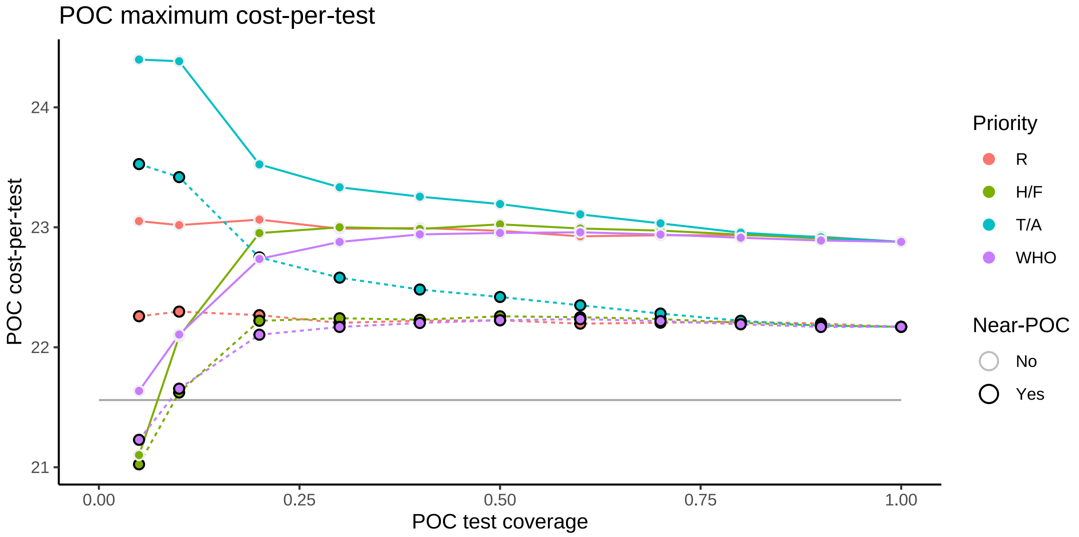


Figure 4.5: Proportion of ART patients on second- or third-line therapies.

Figure 4.6: Average VL suppression rates among patients on ART 12 months after treatment initiation.

Figure 4.7: The maximum cost-per-test at which POC VL monitoring was cost-effective (ICER<500) relative to the baseline scenario of centralized testing alone.



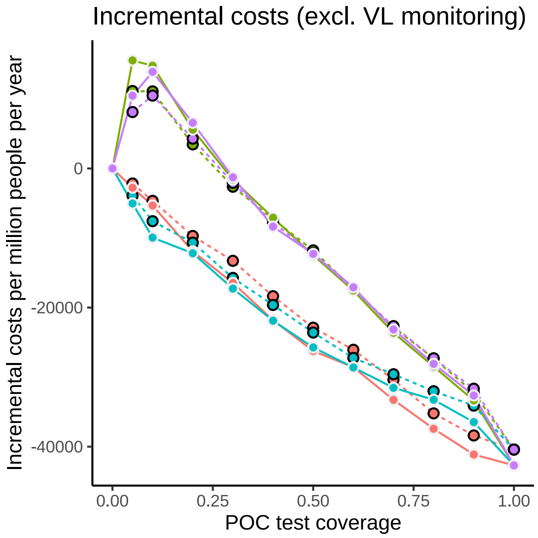


Figure 4.8: Incremental HIV program costs, excluding VL monitoring, per million adults per year.

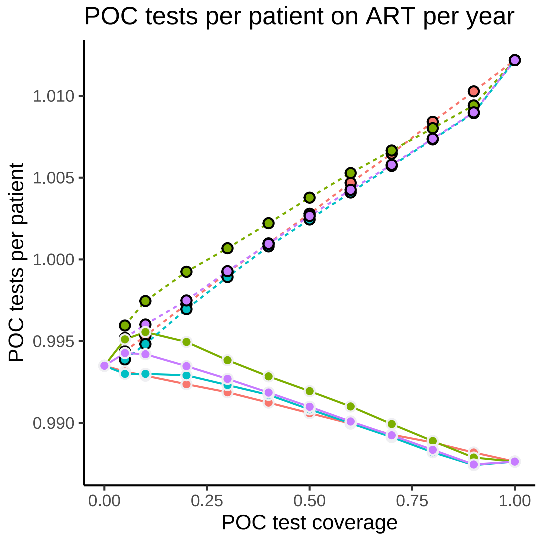


Figure 4.9: Number of VL monitoring samples per patient on ART per year.

in the proportion of POC testing performed. However, prioritizing high-risk patients for access to POC tests provided significantly larger benefits for small amounts of POC testing (5-30%). The steepest change in DALYs was obtained by prioritizing patients with a recent high VL or suspected treatment failure (**H/F**), followed by the WHO recommendations. Prioritizing patients at high risk of transmission and/or poor adherence (**T/A**) was more effective than random test allocation for low levels of POC testing, but performed slightly worse for high levels ($> 80\%$) of POC testing due to the fact that older patients were consistently excluded from POC testing. The impact of near-POC testing followed similar trends in terms of priority strategies and incremental benefits of increased coverage.

Other Outcomes. Figures 4.5 and 4.6 illustrate the impact of POC testing strategies on the number of patients switched to alternative ART regimens, as well as the rate of VL suppression among patients 12 months after ART initiation. In these results, the impact of different priority criteria for targeted POC testing is clearly visible—strategies which prioritized confirmatory tests or patients with suspected treatment failure (**H/F**, **WHO**) had a higher rate of treatment switches, while strategies which prioritized patients recently initiated on ART (**T/A**, **WHO**) had higher VL suppression rates in this population.

Cost-Effectiveness. Based on an incremental cost-effectiveness threshold of $500 per DALY averted, we used the cost and DALY outputs from the simulation model to calculate the maximum cost-per-test at which POC VL monitoring would be cost-effective[*] (Figure 4.7). Full POC testing was cost-effective at an incremental cost of approximately $1.5 more than centralized tests, while near-POC testing was only cost-effective if the average cost-per-test was around $0.75 higher than the cost of centralized testing. The lower cost threshold for near-POC testing was due to a combination of lower effectiveness, as seen in Figure 4.4, and the higher failure rates of near-POC tests. As illustrated in Figure 4.7, near-POC testing scenarios had higher

[*] assuming that centralized testing costs were $21.56 per sample

total sample volumes per patient than either POC or centralized testing. [†]

When POC tests were allocated randomly, the incremental cost thresholds were approximately the same regardless of the proportion of POC tests (as indicated by the solid red line in Figure 4.7). The cost thresholds were more variable for targeted POC testing strategies, particularly when the proportion of samples tested at POC was relatively small (i.e., when the majority of POC tests are allocated to high-risk patients). The **H/F** and **WHO** strategies had significantly lower cost thresholds, while the **T/A** strategy had significantly higher cost thresholds.

The cost thresholds for different priority criteria were linked to both the total volume of samples collected (Figure 4.7), as well as the follow-up care provided as a result of POC tests. As illustrated in Figure 4.5, the **H/F** and **WHO** strategies resulted in a larger number of patients switching to next-line therapies, which were substantially more expensive than standard first-line ART regimens. These strategies were also more likely to identify patients requiring repeated tests, resulting in higher total sample volumes (Figure 4.7). Due to these additional treatment costs, the cost-effectiveness threshold for these priority strategies is lower. By contrast, the **T/A** criteria resulted in fewer patients on next-line therapies and fewer total tests, allowing higher cost thresholds for POC testing.

4.3.3 Optimal Deployment of POC Instruments

In our analysis of strategies for 20% national POC coverage we focused on the impact of three high-level operational decisions: whether to allow individual facilities to use a combination of POC and centralized testing (vs. one or the other), how to prioritize patients for access to limited POC capacity (if both POC and centralized options are available), and whether to allow near-POC testing (subject to the constraints described in Section 4.2.3.5). Figure 4.10 shows the incremental costs and DALYs for the optimal allocation of POC devices for different combinations of these high-level

[†] Although the failure rates in near-POC testing were lower than those of centralized tests, near-POC tests were not subject to the same delays. This allowed repeated or confirmatory tests to be performed in a shorter space of time, leading to higher average sample volumes.

Table 4.6: A summary of optimal POC instrument allocations to transition approximately 20% of national sample volumes to POC or near-POC testing.

	Priority Strategy	Coverage		Facilities		Access		POC Capacity	
		POC	NPOC	POC	NPOC	POC	NPOC	Total	Utilization
Without Near-POC	AoN	20	0	51	0	20	0	1192	55
	Random	20	0	66	0	26	0	1012	64
	WHO	20	0	80	0	43	0	948	69
	H/F	20	0	96	0	53	0	788	83
	T/A	20	0	88	0	50	0	800	81
With Near-POC	AoN	14	6	47	71	14	6	1232	53
	Random	16	4	55	119	20	11	1036	63
	WHO	14	6	68	490	40	51	844	77
	H/F	13	7	65	483	41	56	784	84
	T/A	13	7	61	484	39	57	804	81

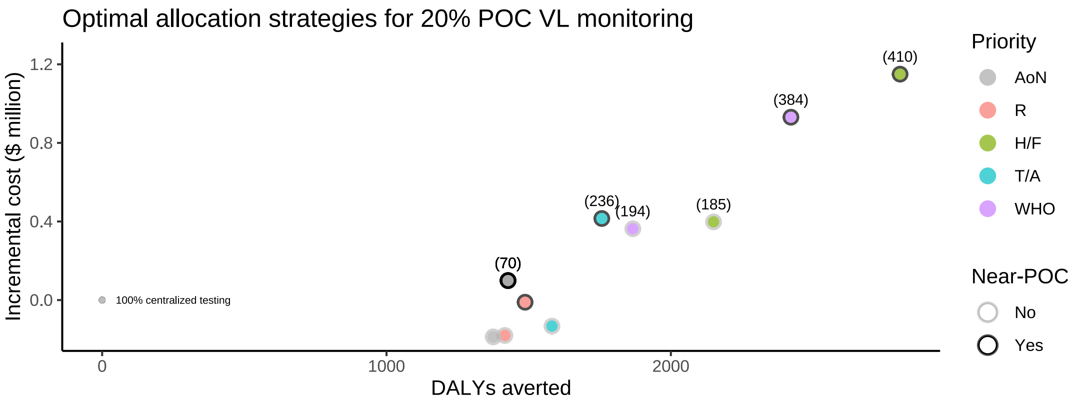


Figure 4.10: Estimated incremental costs and DALYs averted over a 5-year horizon for 20% POC VL monitoring. The ICER for each solution is labeled above the corresponding point.

operational policies, while Table 4.6 provides a summary of the capacity allocations and utilization.

Optimal Policy. The best solution obtained (in terms of DALYs averted) used a combination of POC and near-POC testing and the **H/F** priority strategy. Under this policy, the optimal allocation of POC instruments included 65 POC devices with a total daily capacity of 784 tests and an 84% utilization rate. Near-POC testing was available in a further 483 healthcare facilities. Approximately 6% of national volumes were tested near-POC, and 14% at POC. At a national level, over 97% of ART patients were treated a facility with access to either POC testing (41%) or near-POC testing (56%).

Cost-Effectiveness. As illustrated in Figure 4.10, both the choice of priority strategy and whether to use near-POC testing had a significant effect on the total costs and DALYs. All policies had an ICER of less than $500 due to the constraints in the optimization model, although policies with a larger overall impact on DALYs generally had a higher ICER (primarily due to the increased costs associated with switching additional patients to second-line therapies). Policies without prioritization (all-or-nothing and randomized POC testing) had a negative ICER, indicating that the estimated cost of POC instruments and testing was offset by other cost savings over a 5-year period (i.e., fewer lost or failed tests, less expenditure on centralized testing, and lower overall treatment costs).

Priority Strategies. In terms of total DALYs averted, the relative performance of different priority strategies was similar to the simulation results. The **H/F** priority criteria had the largest impact, followed by the **WHO** criteria and the **T/A** criteria. All three priority strategies had a significantly higher impact than policies which did not include targeted POC testing. As shown in Table 4.6, policies with targeted POC testing generally had significantly higher rates of access to POC testing (40–53%) than those that allocated tests randomly or uniformly (17–26%).

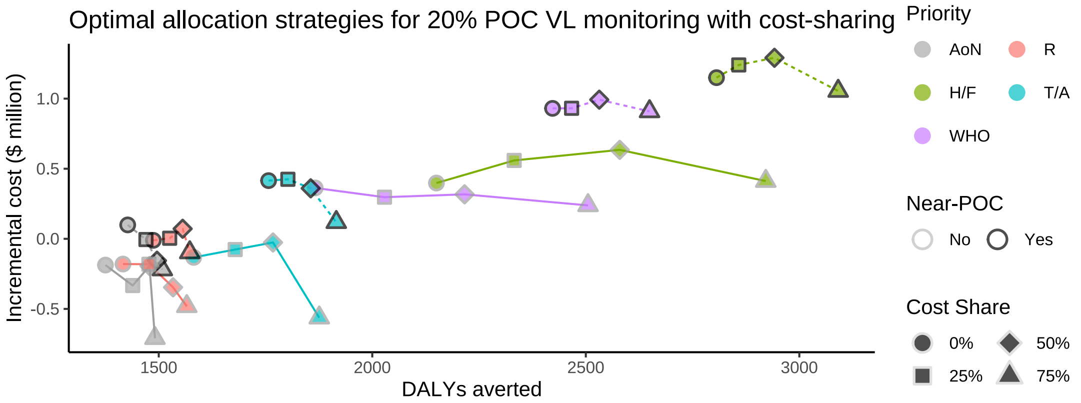


Figure 4.11: A summary of optimal POC allocation strategies with different levels of cost-sharing.

Near-POC Testing. Allowing near-POC testing did not result in significantly higher DALYs averted in the all-or-nothing scenario or when POC tests were allocated randomly, but did have a significant impact when combined with targeted POC testing. For policies that included near-POC testing, the total proportion of near-POC samples was relatively low (3–7%) compared to POC samples (13–16%), but the number of facilities with access to near-POC testing was significantly higher. In scenarios with targeted POC testing, around 70% of all healthcare facilities offered near-POC testing and over 50% of ART patients had access to near-POC testing.

Cost-Sharing. We repeated the analysis described above for scenarios in which the cost and capacity of POC instruments could be shared with other disease programs[‡]. As illustrated in Figure 4.11, policies that allowed cost-sharing were able to achieve significantly higher DALYs and/or lower costs across all priority strategies. The impact of cost-sharing was particularly significant in scenarios where near-POC testing was not permitted. For example, allowing up to 50%–75% cost-sharing without near-POC testing produced similar DALYs and lower costs than policies that allowed near-POC testing without cost-sharing.

Higher POC Coverage. We conducted additional analysis to compare the optimal POC deployment strategies for different national coverage targets ranging from 10% to

[‡] except for instruments used for near-POC testing

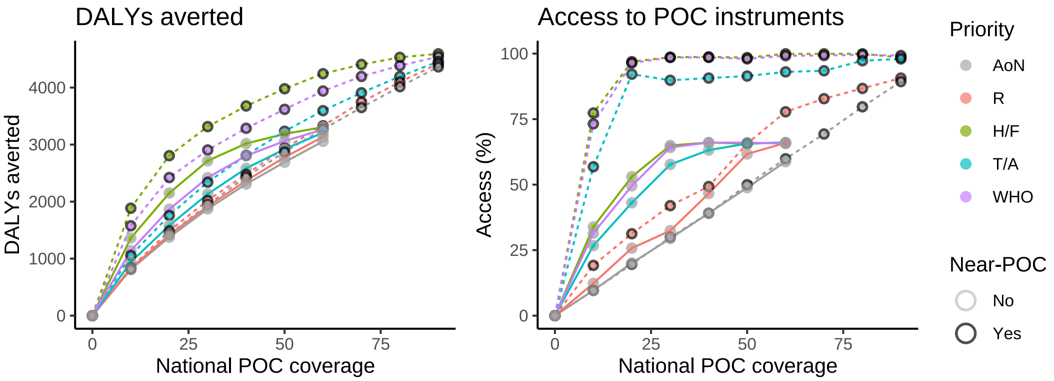


Figure 4.12: A summary of the optimal POC allocation strategies for a range of national coverage targets.

90% of national volumes. As illustrated in Figure 4.12, higher coverage rates resulted in a larger impact (in terms of DALYs averted). Policies that included targeted POC testing were consistently more effective and had higher rates of access to POC testing than randomized or all-or-nothing policies, although the incremental benefit diminished as the total amount of POC testing increased.

In scenarios without near-POC testing, it was not feasible to provide cost-effective POC coverage for more than 60% of national sample volumes unless the cost-effectiveness constraints were relaxed for smaller facilities. We conducted a sensitivity analysis to compare the effects of national vs. facility-level cost-effectiveness constraints, as well as changes of up to 20% in fixed and variable costs associated with different types of testing (see Appendix C.3.5). Overall, these results confirm that the high cost of POC instruments is a significant barrier to cost-effective POC testing in small healthcare facilities.

4.4 Discussion

4.4.1 Cost of POC Testing

The average POC testing costs reported in Table 4.5 are similar to cost estimates in other studies on POC testing in various countries in sub-Saharan Africa (see Table 4.7). However, we found significant variability in POC costs across different healthcare facilities, and the average costs are heavily skewed towards large hospitals which processed high volumes of samples.

In general, there were two important factors that impacted the cost of POC testing at each healthcare facility: the average demand for VL monitoring, and the peak demand for testing. Cost estimates in the existing literature generally focus the former, using average sample volumes to estimate capacity requirements and utilization. Similar to previous studies, our analysis indicates that POC testing is substantially more expensive at low-volume sites due to the high cost and low utilization of POC instruments [80, 150, 125]. The most effective way to compensate for low sample volumes was through cost and capacity sharing with other disease programs, which improved utilization and lowered the fixed costs. Similar results have been found for POC VL monitoring in Kenya [37].

To the best of our knowledge, our study is the first to explicitly model the relationship between variability in daily sample volumes and POC costs and utilization. The higher capacity and lower utilization rates for POC-only strategies in Table 4.5 indicate that meeting the peak demand for testing required substantially more capacity than scenarios without this constraint. These results suggest that POC cost-estimates based on average sample volumes alone are likely to either underestimate the amount of capacity required (and therefore the fixed costs), or over-estimate the coverage and utilization rates for POC instruments. The latter is particularly problematic in settings where it is necessary for patients to receive their results on the same day that the sample is collected, or in scenarios where facilities are using 100% POC testing (as it may result in significant backlogs for VL monitoring).

Our analysis included two strategies to mitigate the effect of variability in demand for VL monitoring—referring excess samples for external testing (i.e., centralized testing), and expanding clinic hours to offer VL testing every day of the week. Both of these strategies would facilitate the use of smaller, less expensive POC instruments (Figure 4.3) and a 2-3 fold increase in utilization of POC capacity (Table 4.5). Based on these findings, it is reasonable to assume that any implementation of POC VL monitoring should include at least one, if not both, of these strategies.

Our analysis suggests that the majority of healthcare facilities in Malawi have insufficient demand for VL monitoring to offset the high fixed costs of POC instruments. However, it may be feasible to conduct POC VL monitoring at comparable costs to centralized testing in large facilities such as district and central hospitals, which generate a significant proportion of the national sample volumes.

From a practical perspective, targeting high-volume facilities for the introduction of POC VL monitoring is likely to have logistical and operational advantages. Large hospitals generally have established laboratory infrastructure and staff, and may already perform some types of POC testing (for example, [77] describes a near-POC pilot at two large facilities in Lilongwe). Large facilities are also more likely to be suitable for cost-sharing arrangements and expanded clinic hours, which may help to lower costs and increase utilization.

On the other hand, focusing on high-volume facilities for POC testing would likely widen the existing disparities in the quality of diagnostic services available to people living with HIV. Facilities with low sample volumes are disproportionately located in rural areas and are likely to have more limited access to diagnostic services and experience longer turnaround times for centralized testing [135, 81]. Although POC testing is often seen as a potential solution to improve diagnostic services in these settings, the POC instruments available for purchase are unsuitable for locations with low demand in resource-limited settings. It is therefore likely that some degree of centralized and/or near-POC testing will be required in order to ensure that VL monitoring is accessible to all patients.

4.4.2 Impact of POC Testing

The results obtained from the HIV Synthesis simulations indicate that replacing centralized VL monitoring with POC testing in Malawi would have a relatively modest impact of approximately 95 DALYs averted per million adults per year within the first five years. These results are similar to estimates obtained using earlier versions of the model, which also noted a link between reduced failure rates for POC testing and lower VL monitoring costs, as well as increased treatment costs associated with

Estimated costs		Country/	
Central	POC	Study	Assumptions
$25.98	$25.39	South Africa [196]	600 samples per year, GeneXpert IV operating for $>$ 10 hours per day (24 samples), 365 days per year.
$28.62	$24.92–$35.46	Malawi [77]	GeneXpert IV. POC and near-POC study at 2 high-volume clinics. Lower cost estimates assume 75% utilization.
$17.22	$23.23	Zambia [81]	GenXpert Omni (discontinued), 50% utilization (15 tests/week). Centralized costs exclude transportation.
	$25.37–$27.70	Zimbabwe [142]	GeneXpert IV, daily capacity of 16 tests and 60-90% utilization
$25.65	$17.07–$54.93	Kenya [37]	GeneXpert IV, 20-500 VL samples per month, capacity sharing up to 75% with TB testing.

Table 4.7: A summary of estimated costs for POC and centralized testing in recent literature.

faster transitions to next-line therapies [169, 173].

The simulated impact of POC testing on VL suppression rates is relatively small compared to impact observed in the STREAM trial in South Africa [62], which compared outcomes between group of patients who received POC VL monitoring and a control group who received centralized testing. This study reported a 10% increase in VL suppression rates (93% vs. 83%) for patients who had POC VL monitoring vs. centralized testing, whereas the simulation results showed an increase of only 1.6% in VL suppression rates (91.2% vs. 92.8%) in a similar patient group. The significantly higher baseline VL suppression rates in the simulation are primarily due to the introduction of dolutegravir-based ART regimens in 2019, which are more effective than the NNRTI-based regimens in use during the STREAM trial in 2017 [101]. This discrepancy highlights the importance of evaluating potential changes to VL monitoring policies within the broader context of other HIV treatment guidelines, which may evolve rapidly and inconsistently in different settings.

Near-POC Testing. The simulation results indicate that near-POC testing was around 20% less effective than POC testing in terms of DALYs averted, but still offered significant benefits relative to centralized testing. As illustrated in Figure C.5, the reduced impact of near-POC testing was closely correlated with the assumption that 20% of near-POC results would be subject to similar delays as centralized results, and was very sensitive to changes in this assumption (see Figure C.5). In practice, the relative impact of near-POC testing is likely to vary in different contexts depending on the efficiency and reliability of the implemented systems. For example, [29] reported substantial differences in the outcomes of near-POC testing in several countries in sub-Saharan Africa. Despite this variability, the empirical observations reported in these studies were generally consistent with the assumption that the performance of near-POC testing is likely to be somewhere between centralized and POC testing—near-POC tests generally had significantly shorter follow-up times than centralized results, but same-day results were quite rare ($< 8\%$).

Priority Strategies. The simulated **H/F** and **T/A** priority criteria highlight two substantially different approaches to targeted testing: the **H/F** criteria prioritized individual patients who were known to be experiencing problems on their current treatments based on their recent medical history, while the **T/A** criteria prioritized groups of patients who were statistically more likely to have poor adherence, or to transmit the virus to other individuals if their VL was elevated.

In the former group, appropriate follow-up care often required switching treatments, and delayed follow-up care was more likely to result in poor health individual outcomes such as opportunistic infections and death. As a result, providing faster VL testing to these patients was both more expensive (in terms of overall treatment costs) and more effective (in terms of DALYs averted) than the general population. By contrast, patients meeting the **T/A** priority criteria were less likely to need follow-up care, and if they did, it was more likely to be a relatively inexpensive intervention such as adherence counseling and a follow-up test. This contrast indicates that POC VL monitoring is likely to have different impacts on different groups of patients, and so the choice of priority criteria for targeted POC testing should be carefully evaluated in the context of the broader goals of HIV treatment programs.

Our analysis of different priority criteria has some limitations related to the structure and assumptions of the simulation model. Since the model does not include children under the age of 15 or model the effects of ART during breastfeeding, the simulation results may underestimate the impact of prioritizing adolescents and breastfeeding mothers for POC testing. The 5-year timeline used in our analysis also places greater emphasis on short- and medium-term outcomes such as opportunistic infections and death, and may underestimate the potential benefits of reduced transmission and improved VL monitoring for young, healthy individuals.

Cost-Effectiveness. The simulation results indicate that while POC testing does offer benefits relative to centralized testing, these benefits are not large enough to justify significantly higher costs for POC tests. Notably, the maximum POC cost thresholds calculated from the simulation output (Figure 4.7) are generally lower than the average POC cost estimates for healthcare facilities in Malawi (Table 4.5), except in scenarios where at least 50% of the cost of POC instruments is shared by other disease programs. It is therefore unlikely to be cost-effective to transition to 100% POC VL monitoring, or to implement POC testing uniformly across all healthcare facilities.

4.4.3 Optimal Deployment of POC Instruments

The results in Section 4.3.3 demonstrate that optimizing the deployment and usage of POC instruments could have a significant impact on both the cost and effectiveness of POC VL monitoring. Our key finding is that the effectiveness of limited POC testing is primarily determined by how successfully these tests are targeted towards *patients who are likely to experience a material difference in treatment* due to faster and/or more reliable VL monitoring.

Priority strategies. To the best of our knowledge, our study is the first to optimize the costs and impact associated with different strategies for targeted POC testing within individual facilities. POC cost models in the literature generally focus on *which facilities* should perform POC testing, rather than which patients should receive POC testing. For example, [150] developed a geospatial optimization model to identify which hard-to-reach facilities in Zambia could use POC or near-POC testing as an alternative to centralized testing, and [80] found that targeted allocation of POC devices to facilities with lower VL suppression rates was more cost-effective than a complete transition to POC at all facilities in South Africa. Although the criteria considered in the latter model are somewhat similar to the **H/F** strategy in our work, the allocation strategies studied in [80] are most similar to the all-or-nothing scenarios in our analysis, as each facility offers only a single type of testing.

By contrast, clinical studies on POC VL monitoring generally take place in settings where centralized testing is also available, and POC testing is targeted towards certain groups of patients based on specific inclusion criteria. [29] reports a variety of criteria used in near-POC testing studies in sub-Saharan Africa, including prioritization of patients at high risk of elevated VL (children and adolescents, patients with recent high VL or suspected treatment failure) or at high risk of transmission (pregnant and breastfeeding women). The STREAM trial focused on a cohort of patients who had recently started first-line ART, while [77] focused on patients in Malawi who had a recent high VL result, symptoms of advanced disease, or required a follow-up VL after switching to second line therapies. In the latter study, 61% of near-POC tests returned high VL results, indicating that facility staff were able to identify patients likely to require follow-up care with a relatively high degree of success. Given the widespread use of targeted testing in clinical studies, it is reasonable to expect this approach would be feasible in the context of a national POC testing program.

Near-POC Testing. The main value of near-POC testing in our analysis was to expand the number of facilities with cost-effective access to POC instruments, which is similar to observations in [81]. This had a significant impact on DALYs in the targeted testing scenarios, where increased access allowed a greater number of high-risk patients to receive faster VL results. However, allowing near-POC testing had minimal benefits in the randomized/all-or-nothing scenarios, where the impact of testing a small proportion of patients from many different facilities was similar to the impact obtained by testing the same number of patients within a single facility.

From a logistical perspective, near-POC testing is likely to be more challenging to implement due to the large number of facilities involved and the need for sample transportation. To mitigate these concerns, we constructed policies which make use of the existing sample transport operations and infrastructure at district hospitals. It is possible that better results could be achieved by relaxing these constraints and allowing more flexible near-POC clusters such as those described in [81], although this would likely require deployment of additional staff, transportation capacity, and infrastructure upgrades.

Cost-Sharing. Similar to near-POC testing, sharing POC instruments with other types of testing has the potential to create much wider access to cost-effective POC VL monitoring. From an operational perspective, this approach is preferable to near-POC testing because it has faster turnaround times and does not require sample transportation. Our analysis considers sharing up to 75% of POC instrument capacity with other types of samples, which is similar to the scenarios considered in [37]. In practice, this degree of cost sharing is unlikely to be feasible in the locations where it would be most beneficial, i.e., small healthcare facilities with insufficient sample volumes to achieve adequate utilization of even the smallest POC instruments. Due to these limitations, near-POC testing is likely to be a more realistic option for widespread POC access.

Higher POC Coverage. The supplementary analysis conducted for higher POC coverage targets indicates that while POC testing is not cost-effective at many facilities, it may be possible to achieve high levels ($80-90\%$) of national POC coverage through a combination of near-POC testing and/or cost-sharing with other disease programs. Similar to the findings in [150], our results suggest that increasing national POC coverage rates from low ($\leq 20\%$) to moderate levels ($20-80\%$) would likely result in more cost-effective solutions due to economies of scale, but that very high levels of POC coverage ($> 90\%$) would be substantially more expensive due to the inclusion of low-volume facilities.

In practice, a rapid national transition from centralized testing to high volumes of POC testing ($> 80\%$) is unlikely to be logistically feasible due to the high start-up costs and shortage of qualified staff. It is also unclear whether POC testing at this scale would deliver the assumed benefits, or simply replicate the backlogs and operational inefficiencies seen in centralized networks. Even if efficient, large-scale near-POC testing is possible, it is reasonable to expect that the resources required to achieve this may be more effectively used to improve the existing centralized infrastructure and implement complimentary solutions to reduce follow-up times for centralized results.

4.5 Conclusion

In conclusion, our analysis has shown that the high cost of POC instruments is a significant barrier to cost-effective POC testing at healthcare facilities with low demand for VL monitoring. As a result, POC testing is unlikely to be a feasible replacement for centralized testing, especially in the locations where it would have the most substantial logistical benefits (i.e., remote areas with limited access to diagnostic services).

Based on our analysis of various operational strategies for the implementation of limited POC VL monitoring in Malawi, the most significant impact is obtained when healthcare facilities are able to use both POC and centralized testing in a complimentary manner. The most successful policies include appropriate priority criteria for allocating limited POC tests to high-risk patients, as well as operational strategies such as near-POC testing or cost-sharing, which enable cost-effective access to POC instruments at a larger number of facilities.

Chapter 5

Globally Optimized Survival Trees

5.1 Introduction

Survival analysis is a cornerstone of healthcare research and is widely used in the analysis of clinical trials as well as large-scale medical datasets such as Electronic Health Records and insurance claims. Survival analysis methods are required for censored data in which the outcome of interest is generally the time until an event (onset of disease, death, etc.), but the exact time of the event is unknown (censored) for some individuals. When a lower bound for these missing values is known (for example, a patient is known to be alive until at least time t) the data is said to be right-censored.

A common survival analysis technique is Cox proportional hazards regression [53] which models the hazard rate for an event as a linear combination of covariate effects. Although this model is widely used and easily interpreted, its parametric nature makes it unable to identify non-linear effects or interactions between covariates [31].

Recursive partitioning techniques (also referred to as *trees*) are a popular alternative to parametric models. When applied to survival data, survival tree algorithms partition the covariate space into smaller and smaller regions (*nodes*) containing observations with homogeneous survival outcomes. The survival distribution in the final partitions (*leaves*) can be analyzed using a variety of statistical techniques such as Kaplan-Meier curve estimates [102]. Several authors have proposed algorithms for building survival trees using censored datasets [207, 116, 95], many of which have been implemented within recursive partitioning software packages [206, 94].

Most recursive partitioning algorithms generate trees in a top-down, greedy manner, which means that each split is selected in isolation without considering its effect on subsequent splits in the tree [35, 179, 178]. This approach can have a negative impact on the quality of the model, such as unnecessarily increasing complexity or decreasing accuracy, resulting in poor out-of-sample performance.

To address these issues, researchers have proposed the construction of optimal decision trees, leveraging optimization techniques [45, 153, 192, 224, 222]. Such approaches lead to higher quality solutions while providing the flexibility to impose additional constraints on the trees. As the problem of tree construction is NP-complete [114], recovering the optimal partition in high-dimensional dataset poses scalability issues. [22, 24] have proposed an efficient algorithm which uses modern mixed-integer optimization (MIO) techniques to address this weakness. Similar to other optimization-based approaches, this *Optimal Trees* algorithm forms the entire decision tree in a single step, allowing each split to be determined with full knowledge of all other splits. It allows the construction of single decision trees for classification and regression that have performance comparable with state-of-the-art methods such as random forests and gradient boosted trees, without sacrificing the interpretability offered by a single-tree model.

The key contributions of this chapter are:

1. We present *Globally Optimized Survival Trees* (GOST), a new survival trees algorithm that utilizes the *Optimal Trees* framework to generate interpretable trees for censored data.

2. We propose a new accuracy metric that evaluates the fit of Kaplan-Meier curve estimates relative to known survival distributions in simulated datasets. We also demonstrate that this metric is reasonably consistent with the Integrated Brier Score [86], which can be used to evaluate the fit of Kaplan-Meier curves when the true distributions are unknown.

3. We evaluate the performance of our method in both simulated and real-world datasets and demonstrate improved accuracy relative to two existing algorithms.

4. Finally, we provide examples of how the algorithm can be used in real-world settings with censored data. We apply the algorithm to predict the risk of adverse events associated with cardiovascular health in the Framingham Heart Study dataset, and to predict the risk of mortality in the Wisconsin Longitudinal Study and Health and Lifestyle Survey.

The structure of this chapter is as follows. We review existing survival tree algorithms in Section 5.2 and discuss some of the technical challenges associated with building trees for censored data. In Section 5.3, we give an overview of the Optimal Trees algorithm proposed by [22] and we adapt this algorithm for Globally Optimized Survival Trees in Section 5.4. Section 5.5 begins with a discussion of existing survival tree accuracy metrics, followed by the new accuracy metrics that we have introduced to evaluate survival tree models in simulated datasets. Simulation results are presented in Section 5.6 and results for real-world datasets are presented in Section 5.7. We conclude in Section 5.8 with a brief summary of our contributions.

5.2 Review of Survival Trees

Recursive partitioning methods have received a great deal of attention in the literature, the most prominent method being the Classification and Regression Tree (CART) algorithm [35]. Tree-based models are appealing due to their logical, interpretable structure as well as their ability to detect complex interactions between covariates. However, traditional tree algorithms require complete observations of the dependent variable in training data, making them unsuitable for censored data.

Tree algorithms incorporate a splitting rule that selects partitions to add to the tree, and a pruning rule that determines when to stop adding further partitions. Since the 1980s, many authors have proposed splitting and pruning rules for censored data. Splitting rules in survival trees are generally based on either (a) node distance measures that seek to maximize the difference between observations in separate nodes or (b) node purity measures that seek to group similar observation in a single node [240, 137].

Algorithms based on node distance measures compare the two adjacent child nodes that are generated when a parent node is split, retaining the split that produces the greatest difference in the child nodes. Proposed measures of node distance include the two-sample logrank test [47], the likelihood ratio statistic [46] and conditional inference permutation tests [95]. We note that the score function used in Cox regression models also falls into the class of node distance measures, as the partial likelihood statistic is based on a comparison of the relative risk coefficient predicted for each observation.

Dissimilarity-based splitting rules are unsuitable for certain applications (such as the Optimal Trees algorithm) because they do not allow for the assessment of a single node in isolation. We therefore focus on node purity splitting rules for developing the GOST algorithm.

[85] published the first survival tree algorithm with a node purity splitting rule based on Kaplan-Meier estimates. [56] used a splitting rule based on the negative log-likelihood of an exponential model, while [207] proposed using martingale residuals as an estimate of node error. [115] suggested comparing the log-likelihood of a saturated model to the first step of a full likelihood estimation procedure for the proportional hazards model and showed that both the full likelihood and martingale residuals can be calculated efficiently from the Nelson-Aalen cumulative hazard estimator [146, 1]. More recently, [137] proposed a new approach to adjust loss functions for uncensored data based on inverse probability of censoring weights (IPCW).

Most survival tree algorithms make use of cost-complexity pruning to determine the correct tree size, particularly when node purity splitting is used. Cost-complexity pruning selects a tree that minimizes a weighted combination of the total tree error (i.e., the sum of each leaf node error) and tree complexity (the number of leaf nodes), with relative weights determined by cross-validation. A similar split-complexity pruning method was suggested by [116] for node distance measures, using the sum of the split test statistics and the number of splits in the tree. Other proposals include using the

Akaike Information Criterion (AIC) [47] or using a p-value stopping criterion to stop growing the tree when no further significant splits are found [95].

Survival analysis methods have been extended to include other non-linear learners, such as support vector machines, tree ensembles, and neural networks [73, 93, 123]. [34] adapted the CART-based random forest algorithm to survival data, while both [96] and [98] proposed more general methods that generate survival forests from any survival tree algorithm. "Survival forest" algorithms aggregate the results of multiple trees and aim to produce more accurate predictions by avoiding the instability of single-tree models. In addition, the formulation of the SVM problem has been extended in the survival setting with the objective of maximizing the concordance index for comparable pairs of observations [219, 66]. Neural network survival analysis includes various structures, such as feed forward, deep, and recurrent neural networks [27, 187, 72, 82].

Unlike decision trees, these approaches lead to "black-box" models which are not interpretable and provide little information about how they arrive at their predictions [190, 41]. The issue of interpretability has become central to the adoption and implementation of artificial intelligence models over the past several years [79], particularly in application areas like medicine where algorithmic decisions can directly impact patient lives [183, 38].

More interpretable survival analysis methods are often based on linear models such as Cox proportional hazards regression [53]. Various authors have adapted this approach using regularization techniques such as LASSO [208, 157], ridge regression [223], and elastic net [197], which can be used to perform feature selection in large datasets and control the complexity of the models. Although linear models are relatively easy to interpret, their parametric structure can be a significant limitation if the underlying assumptions (for example, proportional hazards) are violated. These models are are also unsuitable for identifying non-linear relationships and interactions in the data.

Single tree models provide a clear answer to this problem as they are able to capture intrinsic non-linear effects and interactions in the data while offering transparency to the user with the full characterization of potential risk profiles [24].

Relatively few survival tree algorithms have been implemented in publicly available, well-documented software. Two user-friendly options are available in R [180] packages: Therneau's algorithm based on martingale residuals is implemented in the rpart package [206] and Hothorn's conditional inference (ctree) algorithm in the party package [94].

5.3 Review of Optimal Predictive Trees

In this section, we briefly review approaches to constructing decision trees, and in particular, we outline the Optimal Trees algorithm. The purpose of this section is to provide a high-level overview of the Optimal Trees framework; interested readers are encouraged to refer to [24] and [64] for more detailed technical information.

Traditionally, decision trees are trained using a greedy heuristic that recursively partitions the feature space using a sequence of locally-optimal splits to construct a tree. This approach is used by methods like CART [35] to find classification and regression trees. The greediness of this approach is also its main drawback—each split in the tree is determined independently without considering the possible impact of future splits in the tree on the quality of the here-and-now decision. This can create difficulties in learning the true underlying patterns in the data and lead to trees that generalize poorly. The most natural way to address this limitation is to consider forming the decision tree in a single step, where each split in the tree is decided with full knowledge of all other splits.

The first efforts in the direction of optimal decision tree construction involved the use of pattern mining techniques to construct a global model [152, 153]. [143] proposed the use of a Boolean satisfiability model for computing small-size decision trees with optimality guarantees ($n < 10^3$ observations) and [224] introduced an alternative binary formulation that employs Integer Linear Programming to render the model size largely independent from the training data size, achieving better scaling performance and shorter running times for datasets with thousands of observations.

[222] recently suggested an even more efficient way to decompose the learning problem with a constraint programming approach. While there are many other algorithms for constructing globally optimal predictive trees described in the literature [20, 199, 87], these methods generally do not scale to datasets of the sizes required by practical applications (i.e., sample size $n > 20,000$ and number of features $p > 100$), and therefore have not displaced greedy heuristics as the approach used in practice.

Optimal Trees is a novel approach for decision tree construction that outperforms many existing decision tree methods [24]. It formulates the decision tree construction problem from the perspective of global optimality using mixed-integer optimization (MIO) and solves this problem with coordinate descent to find optimal or near-optimal solutions in practical run times. These Optimal Trees are often as powerful as state-of-the-art methods like random forests or boosted trees, yet they produce models composed of a single decision tree and are therefore are readily interpretable.

The Optimal Trees framework is a generic approach that tractably and efficiently trains decision trees according to a loss function of the form

$$\min_{T} \quad \texttt{error}(T, D) + \alpha \cdot \texttt{complexity}(T), \tag{5.1}$$

where T is the decision tree being optimized, D is the training data, $\texttt{error}(T, D)$ is a function measuring how well the tree T fits the training data D, $\texttt{complexity}(T)$ is a function penalizing the complexity of the tree (for a tree with splits parallel to the axis, this is simply the number of splits in the tree), and α is the *complexity parameter* that controls the tradeoff between the quality of the fit and the size of the tree. Cross-validation takes places as an internal component of the method.

Optimal Trees is able scale to large datasets (n in the millions, p in the thousands) by using coordinate descent to train the decision trees towards global optimality. When training a tree, the splits in the tree are repeatedly optimized one-at-a-time, finding changes that improve the global objective value in Equation (5.1). To give a high-level overview, the nodes of the tree are visited in a random order and at each node the following modifications are considered:

- If the node is not a leaf, delete the split at that node;
- If the node is not a leaf, find the optimal split to use at that node and update the current split;
- If the node is a leaf, create a new split at that node.

For each of the changes, we calculate the objective value of the modified tree with respect to Equation (5.1). If any of these changes result in an improved objective value, then the modification is accepted. When a modification is accepted or all potential modifications have been dismissed, the algorithm proceeds to visit the nodes of the tree in a random order until no further improvements are found, meaning that this tree is a locally optimal for Equation (5.1). The problem is non-convex, so we repeat the coordinate descent process from various randomly-generated starting decision trees, before selecting the final locally-optimal tree with the lowest overall objective value as the best solution. For a more comprehensive guide to the coordinate descent process, we refer the reader to [24].

Although only one tree model is ultimately selected, information from multiple trees generated during the training process is also used to improve the performance of the algorithm. For example, the Optimal Trees algorithm combines the result of multiple trees to automatically calibrate the complexity parameter (α). [24] present a tailored approach for tuning continuous hyperparameters of the algorithm discretize the range of the parameter, identifying a unique mapping between intervals and the corresponding $\texttt{complexity}(T)$. Thus, during the tuning process only a restricted set of values are tested, avoiding the exploration of values that result in overlapping solutions. To properly measure variable importance in light of the fact that only one of many correlated covariates could make it into a single tree, the Optimal Trees framework calculates a variable importance score in the same way as random forests or boosted trees to measure the importance of variables during the entire training process and not just in the final tree. More detailed explanations of these procedures can be found in [64].

The coordinate descent approach used by Optimal Trees is generic and can be

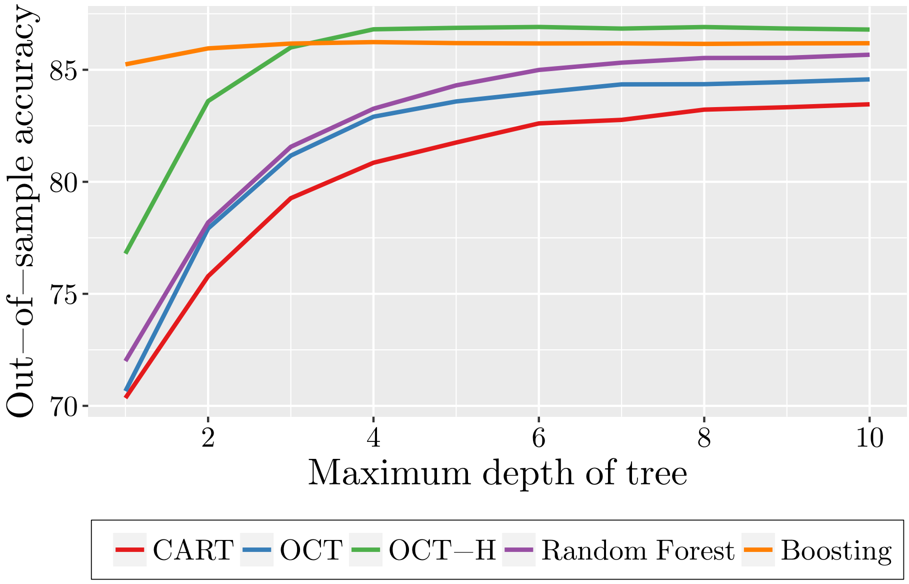


Figure 5.1: Performance of classification methods averaged across 60 real-world datasets. OCT and OCT-H refer to Optimal Classification Trees without and with hyperplane splits, respectively.

applied to optimize a decision tree under any objective function. For example, the Optimal Trees framework can train Optimal Classification Trees (OCT) by setting $\texttt{error}(T, D)$ to be the misclassification error associated with the tree predictions made on the training data. We provide a comparison of performance between various classification methods from [24] in Figure 5.1. This comparison shows the performance of two versions of Optimal Classification Trees: OCT with parallel splits (using one variable in each split); and OCT with hyperplane splits (using a linear combination of variables in each split). These results demonstrate that not only do the Optimal Tree methods significantly outperform CART in producing a single predictive tree, but also that these trees have performance comparable with some of the best classification methods.

5.4 Survival Tree Algorithm

In this section, we adapt the Optimal Trees algorithm described in Section 5.3 for the analysis of censored data. For simplicity, we will use terminology from survival analysis and assume that the outcome of interest is the time until death. We begin with a set of observations $(t_i, \delta_i)_{i=1}^n$ where t_i indicates the time of last observation and δ_i indicates whether the observation was a death ($\delta_i = 1$) or a censoring ($\delta_i = 0$).

Like other tree algorithms, the GOST model requires a target function that determines which splits should be added to the tree. Computational efficiency is an important factor in the choice of target function, since it must be re-evaluated for every potential change to the tree during the optimization procedures. A key requirement for the target function is that the "fit" or error of each node should be evaluated independently of the rest of the tree. In this case, changing a particular split in the tree will only require re-evaluation of the subtree directly below that split, rather than the entire tree.

Due to these computational constraints, splits in the GOST model cannot be evaluated by any methods that require the comparison of two or more nodes within the tree. This requirement restricts the choice of target function to the node purity approaches described in Section 5.2.

The splitting rule implemented in the GOST algorithm is based on the likelihood method proposed by [115]. This splitting rule is derived from a proportional hazards model which assumes that the underlying survival distribution for each observation is given by

$$\mathrm{P}(S_i \leq t) = 1 - e^{-\theta_i \Lambda(t)}, \tag{5.2}$$

where $\Lambda(t)$ is the baseline cumulative hazard function and the coefficients θ_i are the adjustments to the baseline cumulative hazard for each observation.

In a survival tree model we replace $\Lambda(t)$ with an empirical estimate for the cumulative probability of death at each of the observation times. This is known as the Nelson-Aalen estimator [146, 1],

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{j:t_j \geq t_i} 1}. \tag{5.3}$$

Assuming this baseline hazard, the objective of the survival tree model is to optimize the hazard coefficients θ_i. We impose that the tree model uses the same coefficient for all observations contained in a given leaf node in the tree, i.e. $\theta_i = \hat{\theta}_{T(i)}$. These coefficients are determined by maximizing the within-leaf sample likelihood

$$L = \prod_{i=1}^{n} \left(\theta_i \frac{d}{dt} \Lambda(t_i) \right)^{\delta_i} e^{-\theta_i \Lambda(t_i)}, \tag{5.4}$$

to obtain the node coefficients

$$\hat{\theta}_k = \frac{\sum_i \delta_i I_{\{T_i = k\}}}{\sum_i \hat{\Lambda}(t_i) I_{\{T_i = k\}}}. \tag{5.5}$$

To evaluate how well different splits fit the available data we compare the current tree model to a tree with a single coefficient for each observation. We will refer to this as a fully saturated tree, since it has a unique parameter for every observation. The maximum likelihood estimates for these saturated model coefficients are

$$\hat{\theta}_i^{sat} = \frac{\delta_i}{\hat{\Lambda}(t_i)}, \quad i = 1, \dots, n. \tag{5.6}$$

We calculate the prediction error at each node as the difference between the log-likelihood for the fitted node coefficient and the saturated model coefficients at that node:

$$\texttt{error}_k = \sum_{i:T(i)=k} \left(\delta_i \log \left(\frac{\delta_i}{\hat{\Lambda}(t_i)} \right) - \delta_i \log(\hat{\theta}_k) - \delta_i + \hat{\Lambda}(t_i) \hat{\theta}_k \right). \tag{5.7}$$

The overall error function used to optimize the tree is simply the sum of the errors across the leaf nodes of the tree T given the training data D:

$$\texttt{error}(T, D) = \sum_{k \in \text{leaves}(T)} \texttt{error}_k(D). \tag{5.8}$$

We can then apply the Optimal Trees approach to train a tree according to this error function by substituting this expression into the overall loss function (5.1). At each step of the coordinate descent process, we determine new estimates for $\hat{\theta}_k$ for each leaf node k in the tree using (5.5). We then calculate and sum the errors at each node using (5.7) to obtain the total error of the current solution, which is used to guide the coordinate descent and generate trees that minimize the error (5.8).

5.5 Survival Tree Accuracy Metrics

In order to assess the performance of the GOST algorithm, we now introduce a number of accuracy metrics for survival tree models. We will use the notation T^{true} to represent a tree model, where $T_i^{true} = T^{true}(X_i)$ is the leaf node classification of observation i with covariates X_i in the tree T^{true}. We will use the notation T^0 to represent a null model (a tree with no splits and a single node).

5.5.1 Review of Survival Model Metrics

We begin by reviewing existing accuracy metrics for survival models that are commonly used in both the literature as well as practical applications.

1. **Cox Partial Likelihood Score**

   The Cox proportional hazards model [53] is a semi-parametric model that is widely used in survival analysis. The Cox hazard function estimate is

   $$\lambda(t|X_i) = \lambda_0(t)\exp\left(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}\right) = \lambda_0(t)\exp\left(\beta^T X_i\right), \tag{5.9}$$

   where $\lambda_0(t)$ is the baseline hazard function and β is a vector of fitted coefficients. This proportional hazards model does not make any assumptions about the form of $\lambda_0(t)$, and its parameters can be estimated even when the baseline is completely unknown [54]. The coefficients β are estimated by maximizing the partial likelihood function*,

   $$L(\beta) = \prod_{\mathrm{t_i uncensored}} \frac{\exp\left(X_i\beta\right)}{\sum_{t_j \geq t_i} \exp\left(X_j\beta\right)} = \prod_{\mathrm{t_i uncensored}} \frac{\theta_i}{\sum_{t_j \geq t_i} \theta_j}. \tag{5.10}$$

*This definition of the partial likelihood assumes that there are no ties in the data set (i.e., no two subjects have the same event time).

For computational convenience, the Cox model is generally implemented using the log partial likelihood,

$$l(\beta) = \log L(\beta) = \sum_{\mathrm{t_i uncensored}} X_i\beta - \log(\sum_{t_j \geq t_i} \exp{(X_j\beta)}). \tag{5.11}$$

In the context of survival trees, we can find the Cox hazard function associated with a particular tree model by assigning one coefficient to each leaf node in the tree, i.e.,

$$\lambda_T(t) = \lambda_0(t) \exp{(\sum_{k \in T} \beta_k \mathbb{1}(T_i = k))} = \lambda_0(t) \exp{(\beta_{T_i})}. \tag{5.12}$$

We define the Cox Score for a tree model as the maximized log partial likelihood for the associated Cox model, $\max_\beta l(\beta|T)$. To assist with interpretation, we also define the Cox Score Ratio (CSR) as the percentage reduction in the Cox Score for tree T relative to a null model,

$$CSR(T) = 1 - \frac{\max_\beta l(\beta|T)}{\max_\beta l(\beta|T^0)}. \tag{5.13}$$

Due to its widespread use in the context of Cox Regression, the Cox Score is a useful metric for assessing the fit of survival tree models and contrasting the structure of these models with more commonly used linear hazard functions. However, it is important to consider the implications of applying a metric designed for continuous hazard predictions in the context of decision trees, which produce a discrete hazard coefficient for every node. Each additional leaf node in the tree allows an additional degree of freedom in equation (5.12), and increasing the number of nodes in the tree may inflate Cox score even if the overall quality of the model does not improve.

Another significant drawback of the Cox score is its reliance on the proportional hazards assumption (5.2). Although this assumption is commonly used in survival analysis, it may not be appropriate in many applications. This metric should be interpreted with caution when comparing the results of survival tree algorithms that use the proportional hazards model in node splitting rules (such as the GOST algorithm) to other algorithms that rely on non-parametric splitting rules.

2. **The Concordance Statistic**

Applying a ranking approach to survival analysis is an effective way to deal with the skewed distributions of survival times as well as censored of the data. The Concordance Statistic, which is most familiar from logistic regression, is another popular metric that has been adapted to measure goodness-of-fit in survival models [90]. The concordance index is defined as the proportion of all *comparable* pairs of observations in which the model's predictions are *concordant* with the observed outcomes.

Two observations are *comparable* if it is know with certainty that one individual died before the other. This occurs when the actual time of death is observed for both individuals (neither is censored) or when the one individual's death is observed before the other is censored. A comparable pair is *concordant* if the predicted risk (ρ) is higher for the individual that died first, and the pair is discordant if the predicted risk is lower for the individual that died first. Thus, the number of concordant pairs in a sample is given by

$$CC = \sum_{i,j} \mathbb{1}_{(t_i > t_j)} \mathbb{1}(\rho_i < \rho_j)\delta_j, \tag{5.14}$$

and the number of discordant pairs is

$$DC = \sum_{i,j} \mathbb{1}_{(t_i > t_j)} \mathbb{1}_{(\rho_i > \rho_j)}\delta_j, \tag{5.15}$$

where the indices i and j refer to pairs of observations in the sample. Multiplication by the factor δ_j discards pairs of observations that are not comparable because the smaller survival time is censored, i.e., $\delta_j = 0$. These definitions do not include comparable pairs with tied risk predictions, so we denote these pairs as

$$TR = \sum_{i,j} \mathbb{1}(t_i > t_j)\mathbb{1}(\rho_i = \rho_j)\delta_j. \tag{5.16}$$

The number of concordant and discordant pairs is commonly summarized using Harrell's C-index [90],

$$H_C = \frac{CC + 0.5 \times TR}{CC + DC + TR}. \tag{5.17}$$

Harrell's C takes values between 0 and 1, with higher values indicating a better fit. Note that randomly assigned predictions have an expected score of $H_C = 0.5$.

More recently, [218] introduced a modified C-Statistic that weights comparable pairs of observations based on the distribution of censoring times,

$$U_{C_t} = \frac{\sum_{i,j}(\hat{G}(t_j))^{-2}\mathbb{1}_{(t_i>t_j,t_j<t)}\mathbb{1}_{(\rho_i<\rho_j)}\delta_j}{\sum_{i,j}(\hat{G}(t_j))^{-2}(\mathbb{1}_{(t_i>t_j,t_j<t)}\mathbb{1}_{(\rho_i>\rho_j)}\delta_j + \mathbb{1}_{(t_i>t_j,t_j<t)}\mathbb{1}_{(\rho_i<\rho_j)}\delta_j)}, \tag{5.18}$$

where $\hat{G}(\cdot)$ is the Kaplan-Meier estimate for the censoring distribution. Due to these coefficients, U_C converges to a quantity that is independent of the censoring distribution. U_C takes values between 0 and 1, with higher values indicating a better fit.

The above definition of Uno's C-statistic was intended for continuous models, and (5.18) may be very unstable in small trees due to the large number of observations with tied risks which are not counted in either the numerator or denominator. To avoid this, we include these pairs of observations in a similar manner to Harrell's C-statistic, i.e., weighted by 0.5 in the numerator and 1 in the denominator. The resulting concordance statistic is

$$U^*_{C_t} = \frac{\sum_{i,j}(\hat{G}(t_j))^{-2}\mathbb{1}_{(t_i>t_j,t_j<t)}\left(\mathbb{1}_{(\rho_i<\rho_j)} + 0.5 \times \mathbb{1}_{(\rho_i=\rho_j)}\right)\delta_j}{\sum_{i,j}(\hat{G}(t_j))^{-2}(\mathbb{1}_{(t_i>t_j,t_j<t)}\mathbb{1}_{(\rho_i>\rho_j)}\delta_j + \mathbb{1}_{(t_i>t_j,t_j<t)}\mathbb{1}_{(\rho_i\leq\rho_j)}\delta_j)}. \tag{5.19}$$

This modification improves the stability of the concordance statistics but also makes these metrics somewhat less informative in the context of discrete models, since a large number of tied pairs tend to dominate both the numerator and denominator. More generally, concordance statistics do not account for incomparable pairs of observations, which may be problematic when there is significant censoring. The binary definition of concordance fails to account for the magnitude of the difference in predicted risks for comparable observations. As a result, these metrics may be less informative in datasets with significant variations in risk.

Unlike the Cox Score, concordance statistics do not explicitly rely on any parametric assumptions. For proportional hazards models it is natural to define the predicted risk in terms of the hazard coefficients in (5.2), i.e., $\rho_i = \theta_i$. However, it is also possible to contrast the predicted risk of a comparable pair of observations via the predicted survival probabilities, the expected survival times, or any other comparable prediction extracted from the model. In our analysis we evaluate concordance based on the predicted survival probabilities extracted from the Kaplan-Meier curves at each node, i.e., $\rho_i(\tau) = 1 - \hat{S}_i(\tau)$. When comparing the risks of a pair of observations, survival probabilities are evaluated at the time of the first event, $\tau = \min\{t_i, t_j\}$.

3. **Integrated Brier score**

The Brier score metric is commonly used to evaluate classification trees [36]. It was originally developed to verify the accuracy of a probability forecast, primarily for weather forecasting. The most common formula calculates the mean squared prediction error:

$$B = \frac{1}{n}\sum_{i}^{n}(\hat{p}(y_i) - y_i)^2, \tag{5.20}$$

where n is the sample size, $y_i \in \{0, 1\}$ is the outcome of observation i, and $\hat{p}(y_i)$ is the forecast probability of this observed outcome. In the context of survival analysis, the Brier score may be used to evaluate the accuracy of survival predictions at a particular point in time relative to the observed deaths at that time. We will refer to this as the Brier Point Score:

$$BP_\tau = \frac{1}{|\mathcal{I}_\tau|}\sum_{i\in\mathcal{I}_\tau}(\hat{S}_i(\tau) - \mathbb{1}_{(t_i>\tau)})^2, \tag{5.21}$$

$$\text{where } \mathcal{I}_\tau = \{i \in \{1,\ldots,n\}, |t_i \geq \tau \text{ or } \delta_i = 1\}.$$

In this case, $\hat{S}_i(\tau)$ is the predicted survival probability for observation i at time τ and $\mathcal{I}_\tau$ is the set of observations that are known to be alive/dead at time τ. Observations censored before time τ are excluded from this score, as their survival status is unknown.

Applying this version of the Brier score may be useful in applications where the main outcome of interest is survival at a particular time, such as the 1-year survival rates after the onset of a disease. In the experiments that follow, the point-wise Brier Score will be evaluated at the median observation time in each dataset. For easy interpretation, the reported scores are normalized relative to the score for a null model, i.e.

$$BPR_\tau = 1 - \frac{BP_\tau(T)}{BP_\tau(T^0)}. \tag{5.22}$$

The Brier Point score has two significant disadvantages in survival analysis. First, it assesses the predictive accuracy of survival models a single point in time rather than over the entire observation period, which is not well-suited to applications where survival distributions are the outcome of interest. Second, it becomes less informative as the number of censored observations increases, because a greater number of observations are discarded when calculating the score.

[86] have addressed these challenges by proposing an adjusted version of the Brier Score for survival datasets with censored outcomes. Rather than measuring

the accuracy of survival predictions at a single point, this measure aggregates the Brier score over the entire time interval observed in the data. This modified measure is commonly used in the survival literature and has been interchangeably called the Brier Score or the Integrated Brier Score by various authors [184]. In this chapter, we will refer to the metric specific to survival analysis as the Integrated Brier score (IB), defined as

$$IB = \frac{1}{t_{max}} \frac{1}{n} \sum_{i=1}^{n} \int_{0}^{t_i} \frac{(1 - \hat{S}_i(t))^2}{\hat{G}(t)} dt + \delta_i \int_{t_i}^{t_{max}} \frac{(\hat{S}_i(t))^2}{\hat{G}(t_i)} dt. \tag{5.23}$$

The IB score uses Kaplan-Meier estimates for both the survival distribution, $\hat{S}(t)$, and the censoring distribution, $\hat{G}(t)$. In a survival tree model, these estimates are obtained by pooling observations in each node in the tree, i.e., $\hat{S}_i(t) = \hat{S}_{T(i)}(t)$. The IB score is a weighted version of the original Brier Score, with the weights being $1/\hat{G}(t_i)$ if an event occurs before time t_i, and $1/\hat{G}(t)$ if the event occurs after time t. This metric addresses many of the deficiencies identified in the Cox and concordance scores above: it is non-parametric, counts both censored and uncensored observations, and evaluates accuracy of the predicted survival functions over the entire time horizon.

In subsequent sections, we report a normalized version of this metric, the Integrated Brier score ratio (IBR), which compares the sum of the Integrated Brier scores in a given tree to the corresponding Integrated Brier scores in a null tree[†]:

$$IBR = 1 - \frac{IB(T)}{IB(T^0)}. \tag{5.24}$$

Aside from the limitations already discussed, we note that all of the above metrics are subject to noise and often provide contradictory assessments when comparing different tree models. For example, our empirical experiments comparing three candidate models were only able to identify a non-dominated model for about 30% of the instances. In the other 70% of our test cases, none of the three candidate models

[†] [181] calls this *explained residual variation*

scored at least as high as the other models on all metrics. These limitations make it difficult to obtain an unambiguous comparison between the performance of different survival tree algorithms. To address this challenge, we will now introduce a simulation procedure and associated accuracy metrics that are specifically designed to assess survival tree models.

5.5.2 Simulation Accuracy Metrics

A key difficulty in selecting performance metrics for survival tree models is that the definition of "accuracy" can depend on the context in which the model will be used. For example, consider a survival tree that models the relationship between lifestyle factors and age of death. A medical researcher may use such a model to *identify risk factors* associated with early death, while an insurance firm may use this model to *predict mortality risks* for individual clients in order to estimate the volume of life insurance policy pay-outs in the coming years. The medical researcher is primarily interested whether the model has identified important splits, while the insurer is more focused on whether the model can accurately estimate survival distributions.

In subsequent sections we refer to these two properties as *tree recovery* and *prediction accuracy*. We develop metrics to measure these outcomes in simulated datasets with the following structure:

Let $i = 1, \ldots, n$ be a set of observations with independent, identically distributed covariates $\mathbf{X}_i = (X_{ij})_{j=1}^m$. Let T^{true} be a tree model that partitions observations based on these covariates such that $T_i^{true} = T^{true}(\mathbf{X}_i)$ is the index of the leaf node in T^{true} that contains individual i. Let S_i be a random variable representing the survival time of observation i, with distribution $S_i \sim F_{T_i^{true}}(t)$. The survival distribution of each individual is entirely determined by its location in the tree T^{true}, and so we refer to T^{true} as the "true" tree model.

This underlying tree structure provides an unambiguous target against which we can measure the performance of empirical survival tree models. In this context, an empirical survival tree model T has high accuracy if it achieves the following objectives:

1. Tree recovery: the model recovers structure of the true tree ($T(\mathbf{X}_i) = T^{true}(\mathbf{X}_i)$).

2. Prediction accuracy: the model recovers the corresponding survival distributions of the true tree (i.e., $\hat{F}_{T_i}(t) = F_{T_i^{true}}(t)$).

It is important to recognize that these two objectives are not necessarily consistent, particularly in small samples. For example, models with perfect tree recovery may have a small number of observations in each leaf node, leading to noisy survival estimates with low prediction accuracy.

5.5.2.1 Tree Recovery Metrics

We measure the tree recovery of an empirical tree model (T) relative to the true tree (T^{true}) using the following metrics:

1. **Node homogeneity** The node homogeneity statistic measures the proportion of the observations in each node $k \in T$ that have the same true class in T^{true}. This metric is equivalent to the misclassification error and cluster purity metrics which are commonly used in the clustering and tree-based binary classification evaluation contexts respectively [75, 186]. Let $p_{k,l}$ be the proportion of observations in node $k \in T$ that came from class $\ell \in T^{true}$ and let $n_{k,l}$ be the total number of observations at node $k \in T^{true}$ from class $\ell \in C$. Then,

$$NH = \frac{1}{n} \sum_{k \in T} \sum_{l \in T^{true}} n_{k,l} p_{k,l}. \tag{5.25}$$

   A score of $NH = 1$ indicates that each node in the new tree model contains observations from a single class in T^{true}. This does not necessarily mean that the structure of T is identical to T^{true}—for example, a saturated tree with a single observation in each node would have a perfect node homogeneity score (see Figure 5.2). The node homogeneity metric is therefore biased towards larger tree models with few observations in each node.

2. **Class recovery** Class recovery is a measure of how well a new tree model is able to keep similar observations together in the same node, thereby avoiding

unnecessary splits. Class recovery is calculated by counting the proportion of observations from a true class $\ell \in T^{true}$ that are placed in the same node in T. Let $q_{k,l}$ be the proportion of observations from class $\ell \in T^{true}$ that are classified in node $k \in T$ and let $n_{k,l}$ be the total number of observations at node $k \in T$ from class $\ell \in T^{true}$. Then,

$$CR = \frac{1}{n} \sum_{\ell \in T^{true}} \sum_{k \in T} n_{k,l} q_{k,l}. \tag{5.26}$$

This metric is biased towards smaller trees, since a null tree with a single node would have a perfect class recovery score. It is therefore useful to consider both the class recovery and node homogeneity scores simultaneously in order to assess the performance of a tree model (see Figure 5.2 for examples). When used together, these metrics indicate how well the model T reflects the structure of the true model T^{true}.

The node homogeneity and class recovery scores can also be used to compare any two tree models, T^a and T^b. In this case, these metrics should be interpreted as a measure of structural similarity between the two tree models. Note that when T^a and T^b are applied to the same dataset, the node homogeneity for model T^a relative to T^b is equivalent to the class recovery for T^b relative to T^a, and vice versa. The average node homogeneity score for T^a and T^b is therefore equal to the average class recovery score for T^a and T^b. We will refer to this as the *similarity score* for models T^a and T^b.

5.5.2.2 Prediction Accuracy Metric

Our prediction accuracy metric measures how well the non-parametric Kaplan-Meier curves at each leaf in T estimate true the survival distribution of each observation.

1. **Area between curves (ABC)**

   For an observation i with true survival distribution $F_{T_i^{true}}(t)$, suppose that $\hat{S}_{T_i}(t)$ is the Kaplan-Meier estimate at the corresponding node in tree T (see Figure 5.3).

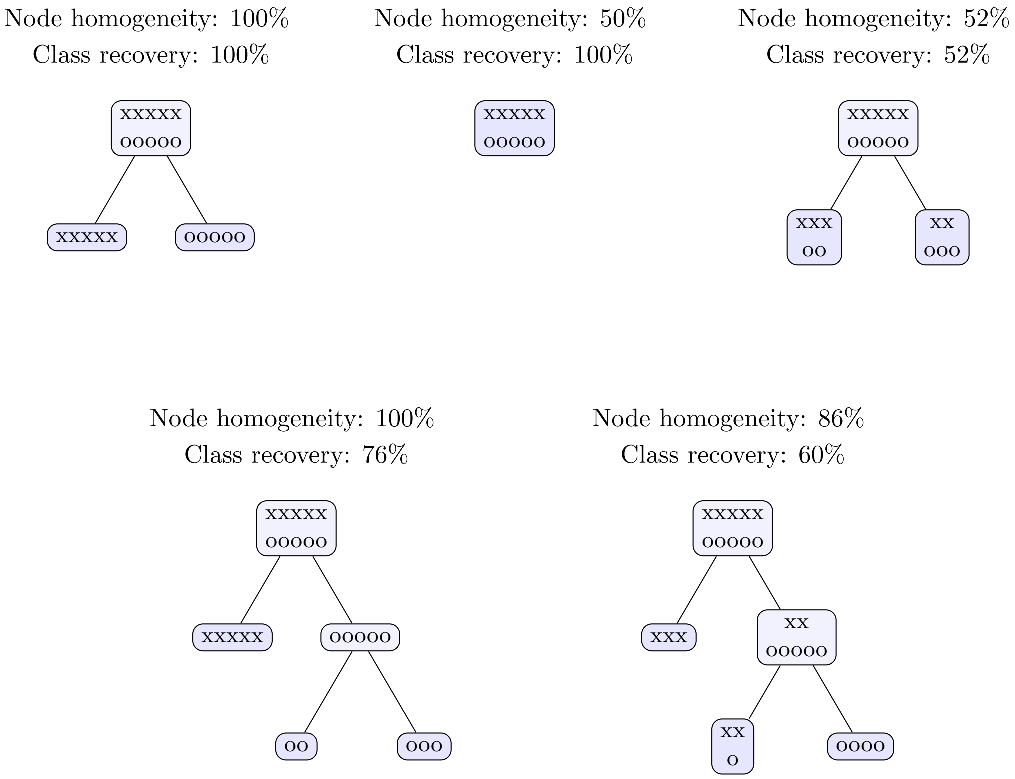


Figure 5.2: Tree recovery metrics for a survival tree with two classes of observations. The top left tree represents the true tree model.

The area between the true survival curve and the tree estimate is given by

$$ABC_i^T = \frac{1}{t_{max}} \int_0^{t_{max}} |1 - F_{T_i^{true}}(t) - \hat{S}_{T_i}(t)| dt. \tag{5.27}$$

To make this metric easier to interpret, we compare the area between curves in a given tree to the score of a null tree with a single node (T^0). The area ratio (AR) is given by

$$AR = 1 - \frac{\sum_i ABC_i^T}{\sum_i ABC_i^{T^0}}. \tag{5.28}$$

Similar to the popular R^2 metric for regression models, the AR indicates how much accuracy is gained by using the Kaplan-Meier estimates generated by the tree relative to the baseline accuracy obtained by using a single estimate for the whole population.

Both the ABC and IBS metrics measure the fit of survival distributions generated at leaf nodes, which are an important component of tree-based survival models. The most important conceptual difference between these metrics is that the IBS compares the estimated survival distributions to events observed in the sampled data (using weights to account for censoring), while the ABC measures accuracy relative to the true survival distributions, which are not affected by censoring or sample size. The ABC cannot be applied in real-world settings where the underlying distributions are unknown, but it provides a simple and intuitive measure of the fit of survival curves in simulation experiments.

5.6 Simulation Results

In this section we evaluate the performance of the Globally Optimized Survival Trees (GOST) algorithm and compare it to two existing survival tree models available in the R packages rpart and ctree. Our tests are performed on simulated datasets with the structure described in Section 5.5.2.

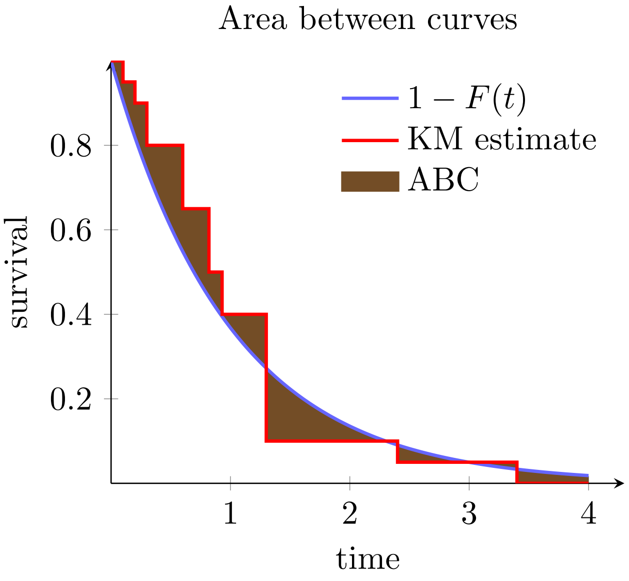


Figure 5.3: An illustration of the area between the true survival distribution and the Kaplan-Meier curve.

5.6.1 Simulation Procedure

The procedure for generating simulated datasets in these experiments is as follows:

1. Randomly generate a sample of 20000 observations with six covariates. The first three covariates are uniformly distributed on the interval $[0, 1]$ and remaining three covariates are discrete uniform random variables with 2, 3 and 5 levels.

2. Generate a random "ground truth" tree model, T^{true}, that partitions the dataset based on these six covariates (see Algorithm 1 in the Appendix).

3. Assign a survival distribution to each leaf node in the tree T^{true} (see Appendix D.1.2 for a list of distributions).

4. Classify observations into node classes $T_i^{true} = C(\mathbf{X}_i)$ according to the ground truth model. Generate a survival time, s_i, for each observation based the survival distribution of its node: $S_i \sim F_{T_i^{true}}(t)$.

5. Generate a censoring time for each observation, $c_i = \kappa(1 - u_i^2)$, where u_i follows a uniform distribution and κ is a non-negative parameter used to control the

proportion of censored individuals.

6. Assign observation times $t_i = \min(s_i, c_i)$. Individuals are marked as censored ($\delta_i = 0$) if $t_i = c_i$.

We used this procedure to generate 1000 datasets based on ground truth trees with a minimum depth of 3 and a maximum depth of 4 (i.e., $2^4 = 16$ leaf nodes). In each dataset, 10000 observations were set aside for testing the tree models. Training datasets of n observations were sampled from the remaining data for sample sizes $n \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$.

In addition to varying the size of the training dataset, we also varied the proportion of censored observations in the data by adjusting the parameter κ. Censoring was applied at nine different levels to generate examples with low censoring (0%, 10%, 20%), moderate censoring (30%, 40%, 50%) and high censoring (60%, 70%, 80%). In total, 63 GOST models were trained for each dataset to test each of the seven training sample sizes at each of the nine censoring levels.

We evaluated the performance of the GOST algorithm relative to two existing survival tree algorithms available in the R packages rpart [206] and ctree [94]. Each of the three algorithms was trained and tested on exactly the same data in each dataset.

Each of the three algorithms tested require two input parameters that control the model size: a maximum tree depth and a complexity/significance parameter that determines which splits are worth keeping in the tree (the interpretation of the ctree significance parameter is different to the complexity parameters in the GOST and rpart algorithms, but it serves a similar function).

Since neither rpart nor ctree have built-in methods for selecting tree parameters, we used a similar 5-fold cross-validation procedure on the training data to select the parameters for each algorithm. We considered tree depths up to three levels greater than the true tree depth and complexity parameter/significance values between 0.001 and 0.1 for the rpart and ctree algorithms (the GOST complexity parameter is automatically selected during training). Equation (5.7) was used as the scoring metric to evaluate out-of-sample performance during cross-validation, and the minimum node size for all algorithms was fixed at 5 observations.

5.6.2 Results

To demonstrate the effect of this cross-validation procedure, we summarize the average size of the models produced by each algorithm in Figure 5.4. We see a clear link between tree size and the number of training observations, indicating the cross-validation procedure is selecting more conservative depth/complexity parameters when relatively little data is available. In larger datasets, the GOST models grow to approximately the same size as the true tree models (6 nodes, on average), while the rpart and ctree models models are slightly larger.

5.6.2.1 Survival Analysis Metrics

Figure 5.5 summarizes the performance of each algorithm in our simulations using the four survival model metrics from Section 5.5. The values displayed in each chart are the average performance statistics across all test datasets.

As expected, the average performance of all three algorithms consistently improves as the size of the training dataset increases. The performance statistics also increase as the proportion of censored observations increases, which seems counter-intuitive (we would expect more censoring to lead to less accurate models). In the case of the Cox partial likelihood and C-statistics, this trend is directly linked to the number of observed deaths, since only observations with observed deaths contribute to the

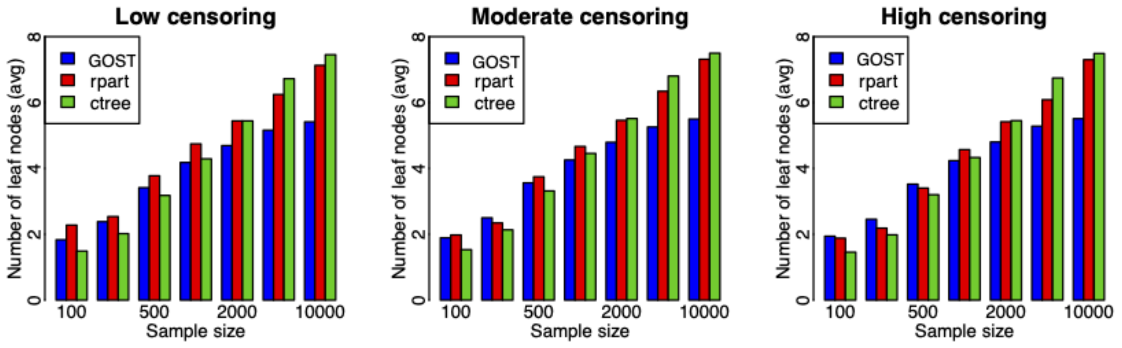


Figure 5.4: The average tree size for models trained on various sample sizes.

Table 5.1: A summary of the average node homogeneity/class recovery scores for simulation experiments.

	Low censoring			Moderate censoring			High censoring		
n	rpart	ctree	GOST	rpart	ctree	GOST	rpart	ctree	GOST
100	38/87	**40**/77	37/**93**	38/90	**40**/78	37/**92**	37/89	**40**/78	37/**90**
200	42/89	**45**/76	43/**91**	42/**90**	**46**/77	45/90	42/**91**	45/78	**45**/90
500	53/84	56/71	**57/88**	55/84	57/70	**59/88**	53/85	56/72	**59/88**
1000	63/82	66/63	**68/89**	65/82	67/63	**70/89**	64/82	66/64	**70/89**
2000	70/81	73/57	**76/89**	72/81	75/57	**78/90**	72/81	74/58	**78/90**
5000	76/80	82/53	**84/91**	77/80	83/53	**85/92**	77/80	82/53	**85/91**
10000	82/79	85/50	**87/91**	84/79	86/51	**89/92**	84/78	86/51	**88/91**

partial likelihood and concordance scores. Similarly, censored observations do not contribute to the Integrated Brier Score after their censoring time.

Each chart also indicates the performance of the true tree model, C, as a point of comparison for the other algorithms. The true tree model performs significantly better than the empirical models trained on smaller datasets, but all three algorithms approach the performance of the true tree for very large sample sizes.

Based on these results, we conclude that the average performance of the GOST algorithm in these simulations is consistently better than either of the other two algorithms. In order to understand why this algorithm is able to generate better models, we now analyse the results of the tree metrics introduced in Section 5.5.2.

5.6.2.2 Tree Recovery

The test set tree recovery metrics for all three algorithms are summarized in Table 5.1 and Figure 5.6. The average node homogeneity/class recovery scores are given side-by-side to allow for a comprehensive assessment of each algorithm's performance. These results confirm that the GOST models perform significantly better than the other two models across all censoring levels.

The node homogeneity scores for all three algorithms increase with larger sample sizes, indicating that the availability of additional data leads to better detection of relevant splits. In large populations, the GOST algorithm selects more efficient splits than the other models and is able to achieve better node homogeneity with fewer

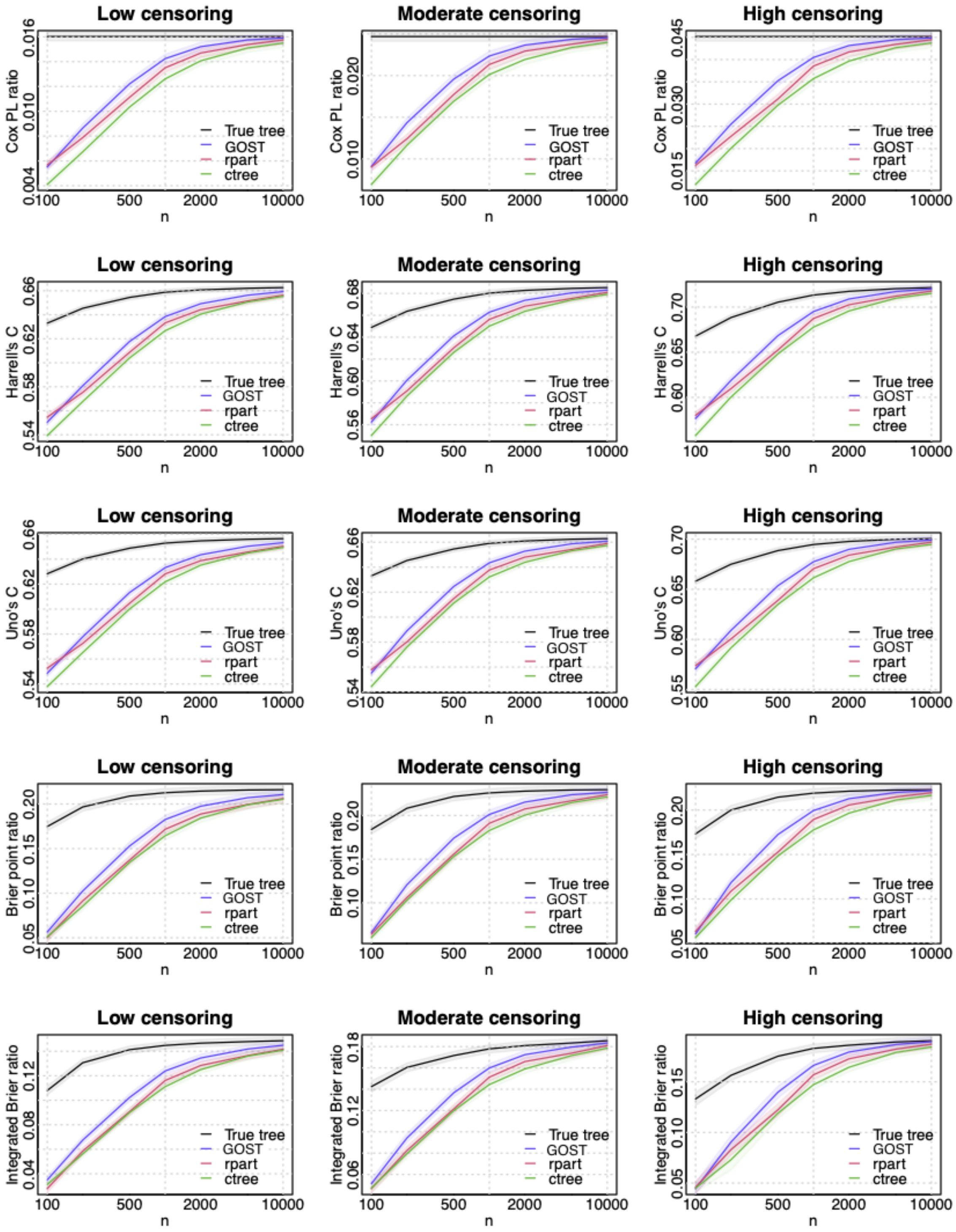


Figure 5.5: A summary of the survival model metrics from simulation experiments. The average test set outcomes for each algorithm are shown in color, while the performance of the true tree model, T^{true}, in indicated in black. Shaded areas indicate 95% confidence intervals.

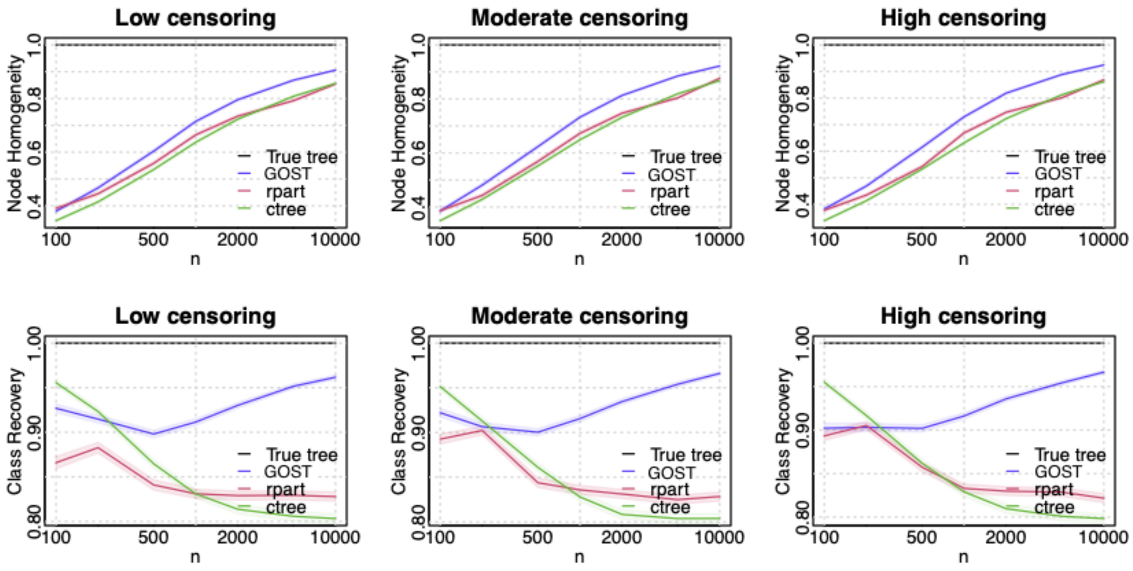


Figure 5.6: A summary of the tree recovery metrics for survival tree algorithms.

splits (recall Figure 5.4—the GOST models trained on large data sets have fewer leaf nodes than the other models, on average).

The relationship between tree size and class recovery rates is somewhat more complicated. In datasets smaller than 500 observations the class recovery rates seem to be closely linked to the tree size: the ctree models have the highest average class recovery for models trained on 100 and 200 observations, and also the smallest number of nodes (see Figure 5.4). However, this trend does not hold in datasets with 500 observations, where GOST models are larger than the ctree models on average, but also have slightly better class recovery. This suggests that tree size is no longer a dominant factor in larger datasets ($n \geq 500$).

In these larger datasets we observe distinct trends in class recovery scores. The GOST class recovery rate increases consistently despite the increases in model size, which means that the GOST models are able to produce more complex trees without overfitting in the training data. By contrast, both of the other algorithms have consistently worse class recovery rates as sample size increases and their models become larger. Based on this trend, neither of these algorithms will reliably converge to the true tree.

Table 5.2: A summary of the average Kaplan-Meier area ratio (AR) scores for simulation experiments.

	Low censoring			Moderate censoring			High censoring		
n	rpart	ctree	GOST	rpart	ctree	GOST	rpart	ctree	GOST
100	6.87	4.79	**9.30**	10.61	7.74	**11.01**	**10.79**	7.76	9.99
200	18.69	16.82	**20.99**	21.93	21.09	**25.25**	24.20	21.24	**26.13**
500	35.03	32.56	**41.17**	40.14	37.12	**47.16**	40.84	38.34	**48.21**
1000	51.27	44.29	**56.44**	57.28	49.68	**61.99**	58.86	51.30	**63.95**
2000	62.76	55.04	**67.97**	68.71	60.30	**73.53**	70.35	61.67	**75.31**
5000	72.62	66.94	**79.45**	77.26	71.63	**83.50**	79.22	72.38	**84.68**
10000	80.06	73.57	**84.41**	84.84	77.44	**87.77**	85.80	77.94	**88.72**

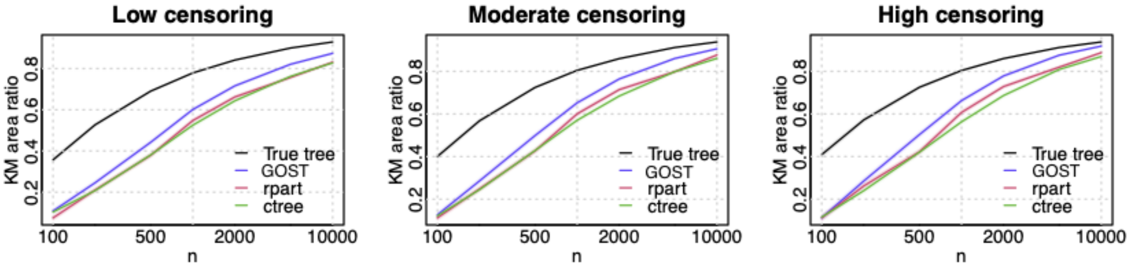


Figure 5.7: A summary of the average Kaplan-Meier Area Ratio results for simulation experiments. The performance of the true tree model is indicated in black.

5.6.2.3 Prediction Accuracy

The test set prediction accuracy metric for each of the three algorithms is summarized in Table 5.2 and Figure 5.7. Overall, the results indicate that sample size plays the most significant role in test set accuracy across all three algorithms. There is also a small increase in accuracy when censoring is increased, which is due to the reduction in the maximum observed time, t_{max}. The GOST results are generally better than the other algorithms across all sample sizes, although the performance gap is relatively small in smaller datasets.

To illustrate the effect of sample size on the accuracy of the Kaplan-Meier estimates, Figure 5.7 also shows the curve accuracy metrics for the true tree, T^{true}. It is immediately apparent that even the true tree models produce poor survival curve estimates in small datasets. Based on these results, it may be necessary to increase the minimum node size to at least 50 observations in applications where Kaplan-Meier curves will be used to summarize survival tree nodes.

5.6.2.4 Comparison of Accuracy Metrics

Table 5.3 shows the correlation between each pair of accuracy metrics used in the simulation experiments. All outcome metrics are positively correlated with the exception of class recovery, which has both weak positive and weak negative correlations with other metrics. These mixed results are due to the different trends in class recovery among the three algorithms – GOST class recovery was highest for trees trained on larger datasets, while the other algorithms had lower class recovery in these instances (see Figure 5.6). Node homogeneity was positively correlated with other metrics, but the correlations were somewhat weaker than average. This reflects the incomplete information captured by this metric – node homogeneity alone does not guarantee a good model, as discussed in Section 5.5.2.1.

Among the other metrics, the highest correlation was observed between the two concordance statistics (0.98), which also had the strongest correlation with most other metrics. There was also high correlation between the two Brier metrics (0.86). The Cox score was most strongly correlated with the concordance statistics (0.87), followed by the Brier statistics (0.77). The Kaplan-Meier area ratio had slightly lower average correlations and was most strongly correlated with the node homogeneity statistic. This is likely due to the fact that both of these metrics are based on the true tree structure, while other metrics reflect how well a model fits the available data.

5.6.2.5 Stability

A frequent criticism of single-tree models is their sensitivity to small changes in the training data. This may be apparent when a tree algorithm produces very different models for different training datasets sampled from the same population. This type of instability is often an indication that the model will not perform well on unseen data.

Given the challenges associated with measuring the test set accuracy for survival tree algorithms, it may be tempting to use stability as a performance metric for these models. Stability is a necessary condition for accuracy in tree models (provided that a tree structure is suitable for the data) but stable models are not necessarily

Table 5.3: Correlation between different accuracy metrics in simulation experiments.

	Cox PL	Harrell's C	Uno's C	Brier point	Integrated Brier	Node Homogeneity	Class Recovery	KM area
Cox PL	1.00	0.87	0.87	0.78	0.77	0.49	-0.03	0.59
Harrell's C	0.87	1.00	0.98	0.90	0.80	0.71	-0.12	0.80
Uno's C	0.87	0.98	1.00	0.87	0.79	0.71	-0.12	0.81
Brier point	0.78	0.90	0.87	1.00	0.86	0.60	0.00	0.71
Integrated Brier	0.77	0.80	0.79	0.86	1.00	0.55	0.02	0.66
Node Homogeneity	0.49	0.71	0.71	0.60	0.55	1.00	-0.03	0.87
Class Recovery	-0.03	-0.12	-0.12	0.00	0.02	-0.03	1.00	0.02
KM area	0.59	0.80	0.81	0.71	0.66	0.87	0.02	1.00

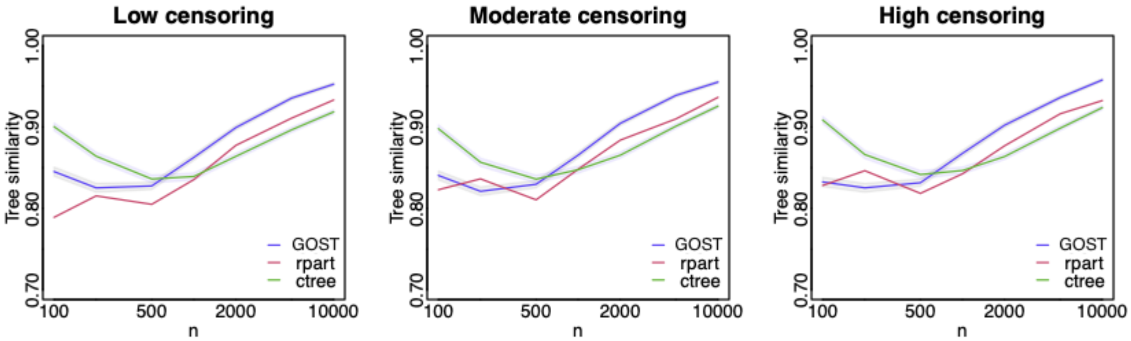


Figure 5.8: A summary of the average similarity scores between pairs of trees trained on mutually exclusive sets of observations.

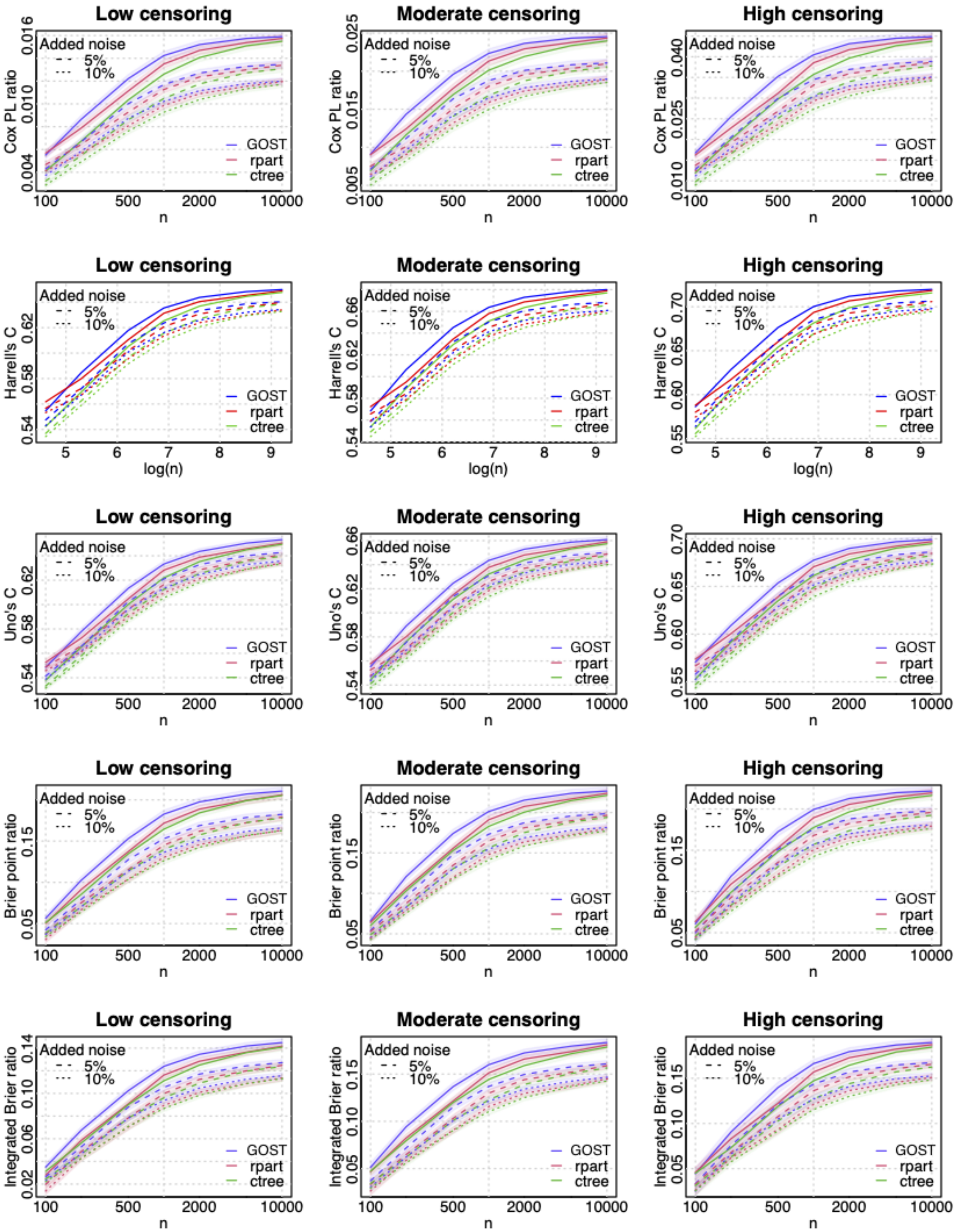


Figure 5.9: A summary of survival tree accuracy metrics for datasets with added noise.

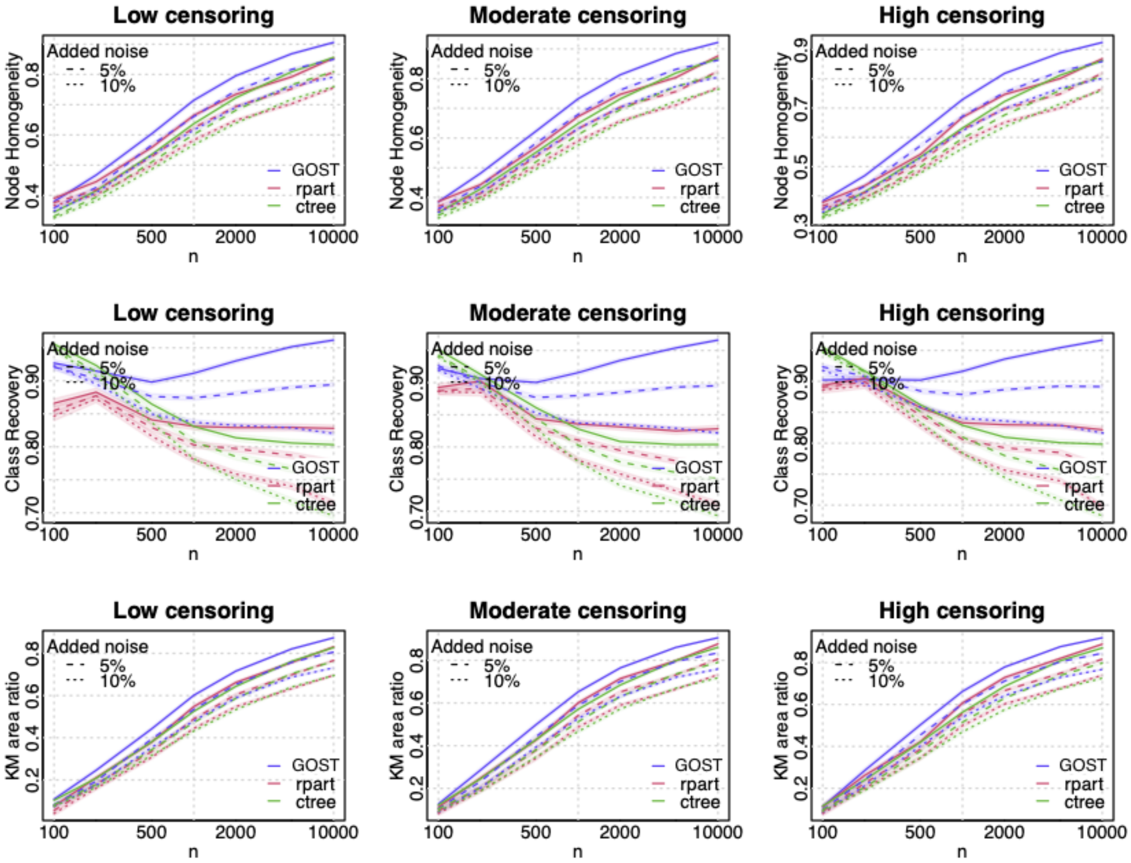


Figure 5.10: A summary of simulation accuracy metrics for datasets with added noise.

accurate. For example, greedy tree models with depth 1 may select the same split for all permutations of the training data, but these models will not be accurate if the data requires a tree of depth 3.

Although stability is not necessarily a good indicator of the quality of a model, it is nevertheless interesting to consider how the stability of globally optimized trees may differ to the stability of greedy trees. Globally optimized trees are theoretically capable of greater stability because they may include splits that are not necessarily locally optimal for a particular training dataset. However, globally optimized trees also consider a significantly larger number of possible tree configurations and therefore have many more opportunities for overfitting on features of a particular training dataset.

We ran two sets of experiments to investigate the stability of the survival tree models in our simulations. In the first set of experiments we used each algorithm to train two models, T^a and T^b, on non-overlapping training datasets of equal size drawn from the same population. We then applied each model to the entire dataset (20000 observations) and used the tree similarity score described in Section 5.5.2.1 to assess the structural similarity between the two models. The average similarity scores for each algorithm are illustrated in Figure 5.8.

These results demonstrate that stability across different training datasets is not a sufficient condition for accuracy: models trained on 100 and 200 observations are both more stable and less accurate than models trained on 500 observations. The ctree algorithm produced the most stable results in smaller datasets due to the smaller model sizes selected during cross-validation. For example, 33.1% of ctree models trained on 100 observations had fewer than 2 splits, compared to 29.5% of the rpart models and 26.5% of the GOST models.

The stability results for larger training datasets ($n > 1000$) are reasonably consistent with the accuracy metrics discussed above, and both stability and accuracy increase with sample size across all three algorithms. The GOST models have the highest average similarity scores in large datasets and the rpart models are slightly more stable than the ctree models.

In the second set of stability experiments we investigated how small perturbations to the covariate values in the training dataset affect the test set accuracy of each model. We added noise to the training data by replacing the original continuous covariate values, x_{ij}, with "noisy" values $\tilde{x}_{ij} = x_{ij} + \epsilon_{ij}$. The initial covariates were uniformly distributed between 0 and 1 and the added noise terms were generated from the following two distributions:

$$\begin{aligned} \epsilon_{ij} &\sim U(-0.05, 0.05) && \text{(5\% noise), and} \\ \epsilon_{ij} &\sim U(-0.1, 0.1) && \text{(10\% noise).} \end{aligned}$$

A similar approach was applied to the categorical variables, which were generated by rounding off continuous values (x_{ij} or $\tilde{x}_{ij}$) to the appropriate thresholds. Note that noise was only added to the observations used for training data; the testing data was unchanged.

The results of these experiments are contrasted with the initial outcomes (without added noise) in Figures 5.9-5.10. The effects of additional noise in the training data are visible in the results of all three algorithms and the drop in accuracy appears to be fairly consistent. Overall, the GOST models maintain the highest scores regardless of noise.

These results indicate that perturbations in the training data affect the GOST and greedy tree algorithms in similar ways. The GOST algorithm's performance is diminished by adding noise to the training data, but its ability to consider a wider range of split configurations does not make it more sensitive to these perturbations. In fact, the GOST algorithm is generally slightly more stable than the greedy algorithms across permutations of the training data because it tends to produce models that are consistently closer to the true tree.

5.6.3 Scaling Performance

We now provide an overview of the computational performance of the GOST algorithm on the synthetic censored datasets. We use the procedure described in Section 5.6.1 to create simulated data varying the number of observations n, the number of features p, and the percentage of censoring. We consider datasets of size $n \in [5000, 10000, 25000, 50000, 100000]$ and $p \in [10, 50, 100]$. We consider three percentages of censoring $[10\%, 50\%, 80\%]$ that correspond to low, moderate, and high censoring respectively. We repeat the experiment for each combination of these parameters on 100 randomized datasets and report the average scaling performance[‡] and the associated 95% confidence intervals. We perform cross validation using grid search to select the best parameters for each model and we report the computational time of the training procedure. Figure 5.11 illustrates our findings.

Across all experiments, the algorithm was able to complete in less than an hour. There was no significant change in the average running time across the different levels of censoring. However, the number of features, p, did have a substantial impact on the computational performance. For $p < 100$, we note that all instances were able to solve within 40 minutes. By contrast, for datasets where the number of covariates is restricted to 10, the average time to solve is less than 25 minutes even when the sample size is 100,000. Increasing the number of observations appears to affect the computational performance in a linear way while the number of features empirically shows an exponential effect.

We present a comparative analysis of the computational performance of the GOST, rpart, and ctree algorithms in Appendix D.1.3. Due to its greedy nature, rpart is able to terminate in less than a minute across all instances. By contrast, we observe that the ctree package requires significantly more time. The latter scales faster than GOST, though it is associated with an exponential rate as the number of observations increases.

[‡] All experiments were conducted on four CPUs of type 2 socket Intel E5-2690 v4 2.6 GHz/35M Cache; 16GB of NUMA enabled memory were used per CPU.

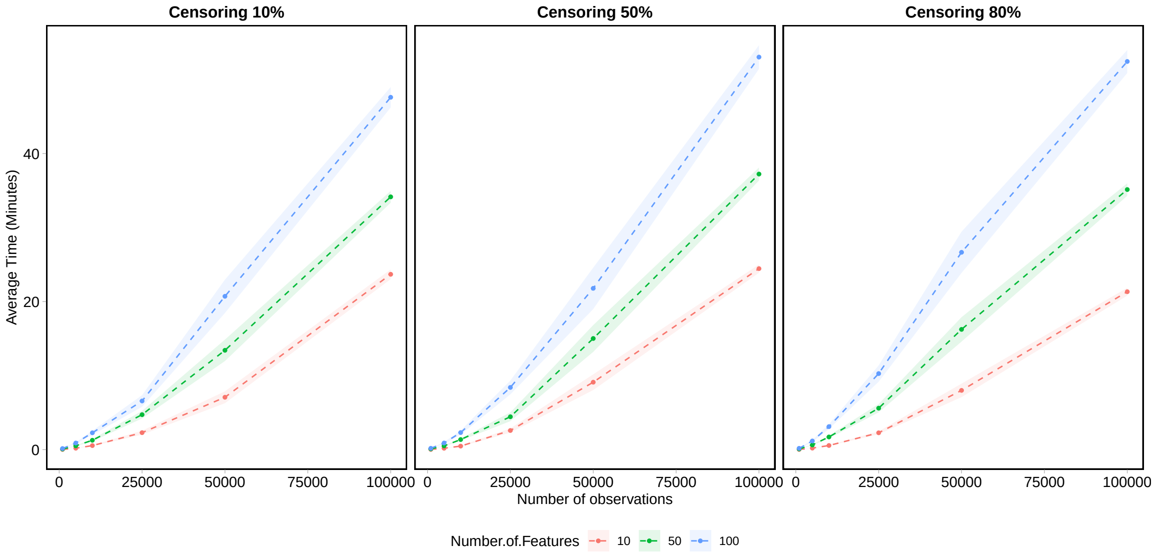


Figure 5.11: Average computational time for GOST tree construction on synthetically generated datasets, with varying numbers of observations n and covariates p. The shaded region corresponds to the 95% confidence intervals.

5.7 Computational Experiments with Censored Data from Longitudinal Studies and Surveys

In this section, we focus on different aspects of algorithmic performance using three widely known longitudinal studies. In Section 5.7.1, we present results from the Wisconsin Longitudinal Study and highlight differences in performance as we vary the mix of categorical and numerical features. In Section 5.7.2, we use data from the Health and Lifestyle survey to compare the algorithms on a large set of features. Finally, in Section 5.7.3, we showcase an application of the algorithm on heart disease using data from the Framingham Heart Study.

The three datasets discussed in this section are typical real-world applications of survival analysis: the outcome of interest is the time to a particular event, and each dataset includes censored outcomes due to individuals lost to follow-up during longitudinal studies. In Appendix D.3, we describe additional experiments in which we simulate different levels of censoring in datasets drawn from the UCI repository [63]. These supplementary results demonstrate the strong performance of the GOST algorithm across a variety of datasets with a range of different sizes and features.

5.7.1 The Wisconsin Longitudinal Study

In 1957, the Wisconsin Longitudinal Study (WLS) randomly sampled 10 317 Wisconsin high school graduates (one-third of all graduates) for a decades-long study, observing them until 2011 [92]. The aim of the study was to understand how factors such as social background, schooling, military service, labor market experiences, family characteristics and events, and social participation, may affect mortality and morbidity, family functioning, and health. We have included in our analysis data from all recorded participants for 518 variables that were collected either from the original respondents or their parents.

We removed from our dataset all features for which more than 50% of the values are missing. We imputed the missing values with the mean of each covariate for numerical features and the mode for categorical and binary variables. In total, we collect 317 categorical, 103 numerical, and 77 binary covariates. In each randomized experiment, we sampled between $[10, 15, 20, 25, 30]$ features from each category. Our goal was to observe the algorithms' performance as we vary the combination of different types of covariates.

Our results show minimal variability in performance as we change the number of numerical and binary features (see Appendix D.2.1). However, all three methods show trends in the average performance scores for different numbers of categorical features, as shown in Figure 5.12. Specifically, both GOST and rpart algorithms show slight decreases in performance with larger feature sets, likely due to overfitting, while the ctree algorithm performs slightly better on larger feature sets.

Overall, GOST clearly outperforms the other methods in terms of the Integrated Brier Score and the Cox PL ratio, and is on par with rpart in both concordance statistics. The ctree algorithm performs poorly relative to the other algorithms across all metrics.

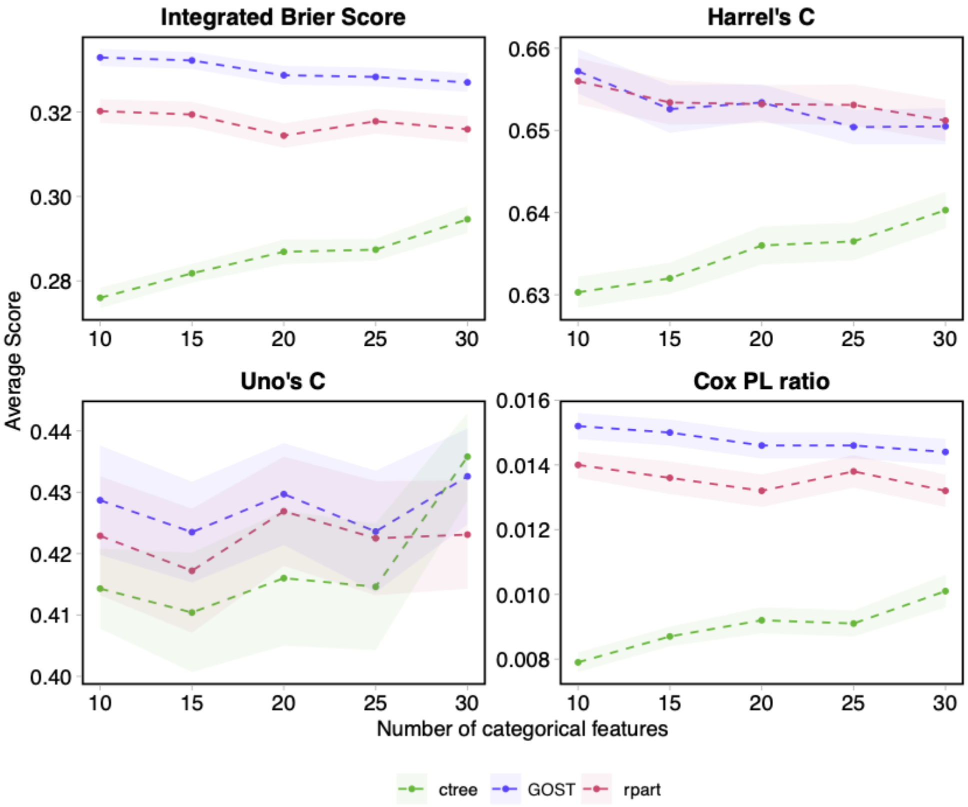


Figure 5.12: Average performance of survival tree models on subsets of features from the WLS dataset with varying numbers of categorical variables. The shaded regions represent 95% confidence intervals across 100 randomized experiments.

5.7.2 The Health and Lifestyle Survey

The first Health and Lifestyle Survey [52] was carried out in 1984-1985 on a random sample of the population of England, Scotland and Wales. Its objective was to help researchers understand the impact of self-reported health, attitudes to health, and beliefs about causes of disease in relation to measurements of health and lifestyle in adults from different parts of Great Britain. In our numerical experiments, the outcome of interest is the age of death of study participants as observed by follow-up studies until 2009. Our dataset includes 9003 individuals and 112 binary features. We conducted 100 randomized experiments to train each tree algorithm.

Table 5.4 outlines the results of our analysis on the HALS dataset. The GOST algorithm outperforms the other methods in all metrics other than the Uno's C metric. Specifically, GOST is associated with an average Integrated Brier Score of 0.6114

Table 5.4: Average scores for GOST, rpart, ctree models on the HALS dataset. For each metric, we report the 95% confidence intervals (in brackets) in 100 randomized experiments.

	IBR		**Cox PL**		**Harrell's C**		**Uno's C**	
GOST	0.6114	(0.608, 0.615)	0.0125	(0.012, 0.013)	0.6211	(0.618, 0.624)	0.3987	(0.391, 0.406)
ctree	0.6056	(0.602, 0.610)	0.0107	(0.010, 0.011)	0.6113	(0.608, 0.615)	0.4098	(0.403, 0.417)
rpart	0.6105	(0.607, 0.614)	0.0124	(0.012, 0.013)	0.6185	(0.615, 0.622)	0.3950	(0.385, 0.405)

compared to 0.6056 and 0.6105 for ctree and rpart respectively. In terms of the Cox PL ratio, GOST offers an 8% improvement over the next best method (rpart) with an average score of 0.0125. With respect to the Harrell's C metric, GOST average Harrell's C metric is 0.6211. ctree and rpart scored 0.6113 and 0.6185 respectively. Contrary to the other measures of performance, ctree achieves the best score in this series of experiments with an average metric of 0.4098 with a 0.0111 margin from GOST. Our findings from this study are in line with the results in Sections 5.6 and supplementary experiments in Appendix D.3.

5.7.3 The Framingham Heart Study

In this section, we focus on the interpretation of the tree models using data from the Framingham Heart Study (FHS). Analysis of the FHS successfully identified the common factors or characteristics that contribute to Coronary Heart Disease (CHD) using the Cox regression model [53]. In our survival tree model, we include all participants in the study from the original cohort (1948-2014) and the offspring cohort (1971-2014) who were diagnosed with CHD. The event of interest in this model is the occurrence of a myocardial infarction or stroke. All 2296 patients were followed for a period of at least 10 years after their first diagnosis of CHD and observations are marked as censored if no event was observed while the patient was under observation.

We applied our algorithm to the primary variables that have been used in the established 10-year Hard Coronary Heart Disease (HCHD) Risk Calculator and the Cardiovascular Risk Calculator [67, 55]. For each participant who was diagnosed with CHD, we include the following covariates in our training dataset: gender, smoking status (smoke), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), use of anti-hypertensive medication (AHT), Body Mass Index (BMI), diabetic status (diabetes). We did not include cholesterol levels in our analysis because these variables are highly correlated with the use of lipid lowering treatment and a high proportion of the sample population did not have sufficient data to account for this interaction.

In Figure 5.13 we illustrate the output of our algorithm on the FHS dataset. Every node of the tree provides the following information:

- The node number.
- Number of observations classified into the node.
- Proportion of the node population which has been censored.
- A plot of survival probability vs. time. In this example, the x-axis represents age and the y-axis gives the Kaplan-Meier estimate for the probability of experiencing no adverse events.
- Color-coded survival curves to describe the different sub-populations. In each node, the blue curves describe the individuals classified into that node.
- In internal (parent) nodes, the orange/green curves describe the sub-populations that are split into the left/right child node. After each split, the sub-population with higher likelihood of survival goes into the left node.
- In leaf nodes, the red curve shows the average survival curve for the entire tree. This facilitates easy comparisons between the survival of a specific node and the rest of the population.

The splits illustrated in Figure 5.13 include known risk factors for heart disease and are consistent with well-established medical guidelines. The algorithm identified a BMI threshold of 25 as the first split (node 1), which is in accordance with the NIH BMI ranges that classify an individual as overweight if his/her BMI is greater than or equal to 25. Multiple splits indicated a higher risk of heart attack or stroke in patients who smoke (nodes 2, 6). The group with the highest risk of an adverse event was overweight patients with diabetes (node 9).

Figures 5.14 and 5.15 illustrate the output of the ctree and rpart algorithms applied to the same FHS population. The rpart model has a single split (BMI), while the ctree model contains the same variables as the GOST output. The Brier scores for each model are 0.0486 (GOST), 0.0249 (rpart) and 0.0467 (ctree).

In this example we can reasonably conclude that the smaller size of the rpart tree impacts its predictive performance. This highlights the important role of cross-validation procedures in selecting an appropriate complexity parameter. In Appendix D.2.2.2 we describe additional experiments which contrast the performance of tree models with uniform size and shape (thus eliminating the effects of parameter selection), and note that the average performance of GOST models is generally better than rpart trees of the same size.

The discrepancy in the Brier scores for the GOST and ctree models is due to slight differences in the threshold and position of certain splits. For example, both methods identify that BMI is the most appropriate variable for the first split, but the BMI threshold differs. The ctree model sets the splitting threshold to 24.117, which is the locally optimal value for the split when building the tree greedily (the same threshold is used in the rpart model). By contrast, the GOST algorithm selects a threshold of 25.031. This example demonstrates how the GOST algorithm's efforts to find a globally optimal solution differ from the results of locally optimal splits.

A second difference between the tree models is the order of the smoking and diabetes splits within the overweight population. The ctree model splits on smoking first, since this split has the most significant p-value of the variables at node 5 in the ctree tree. The algorithm also recognizes that diabetes is a risk factor and incorporates this in the subsequent split. Since greedy approaches like ctree do not reevaluate the spits once they have been decided, the algorithm does not recognize that the overall quality of the tree can be improved by reversing the order of these splits. This discrepancy in two otherwise similar trees highlights the advantages of the more sophisticated optimization conducted by GOST.

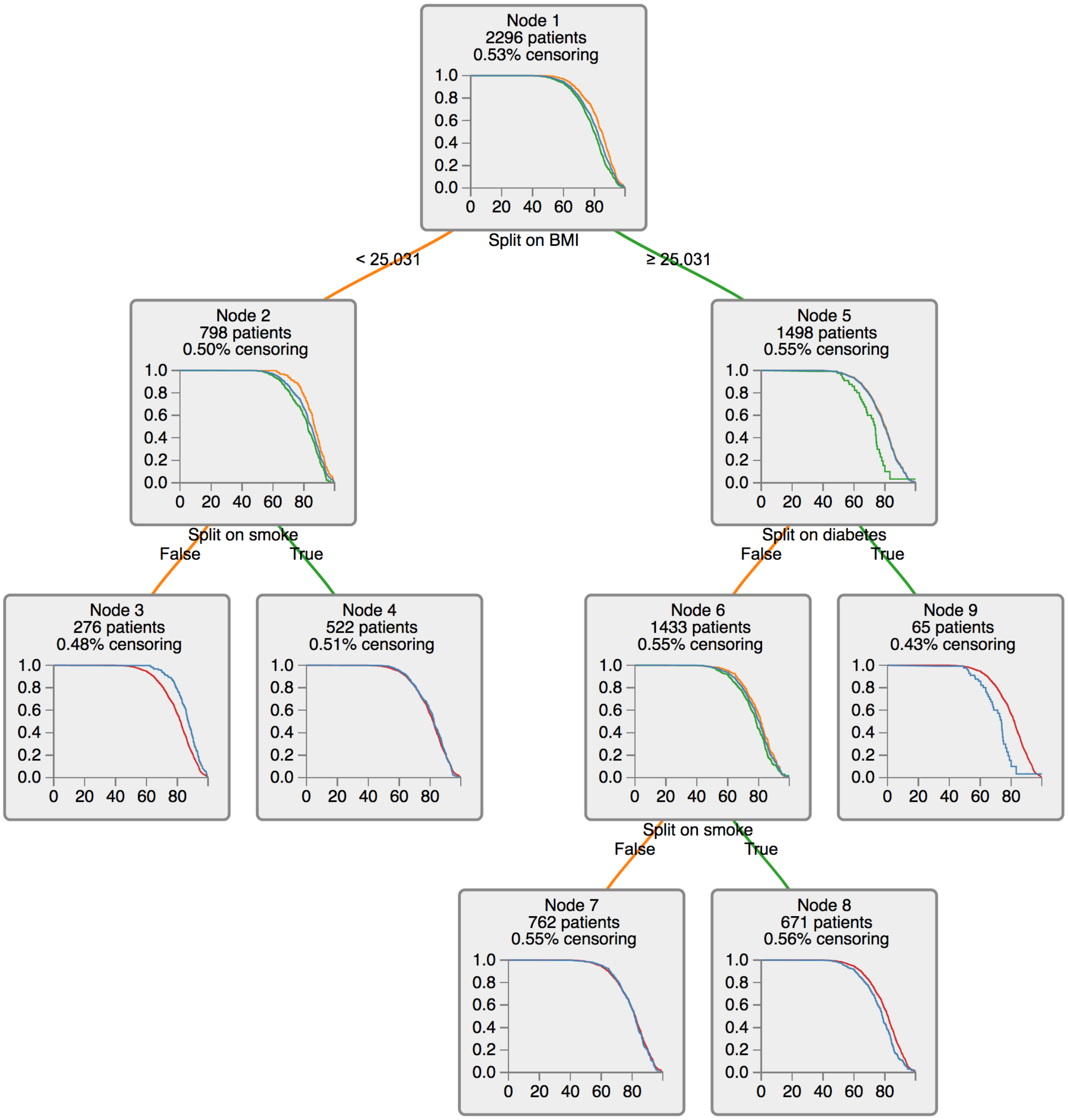


Figure 5.13: An illustration of Globally Optimized Survival Trees for chd patients in the FHS.

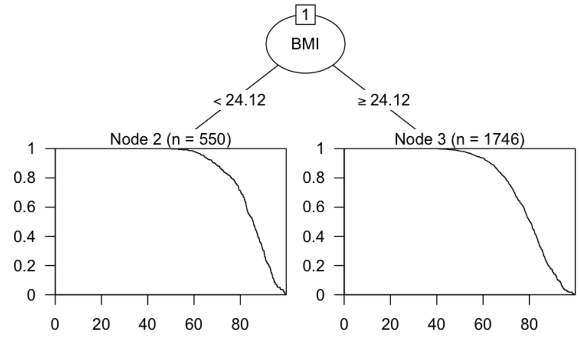


Figure 5.14: Illustration of the rpart output for chd patients in the FHS.

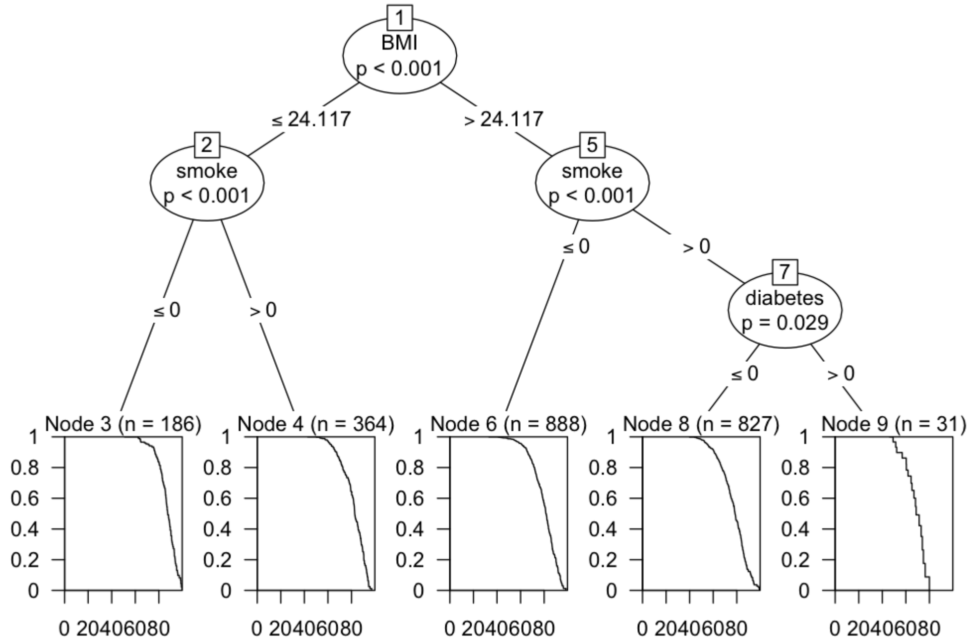


Figure 5.15: Illustration of the ctree output for chd patients in the FHS.

5.8 Conclusion

In this chapter, we have extended the state-of-the-art Optimal Trees framework to generate interpretable models for censored data. We have also introduced a new accuracy metric, the Kaplan-Meier Area Ratio, which provides an effective way to measure the predictive power of survival tree models in simulations.

The Globally Optimized Survival Trees algorithm improves on the performance of existing algorithms in terms of both classification and predictive accuracy. Our results in simulations indicate that the GOST models improve consistently with increasing sample size, whereas existing algorithms are prone to overfitting in larger datasets. This is particularly important, given that the volume of medical data available for research is likely to increase significantly over the coming years.

Appendix A

Summarized Sample Volumes

Table A.1: Summarized sample volumes, by healthcare facility.

	VL	**EID**	**TB**	**Other**	**Total**
Phalombe	**19,354**	**1,405**	**569**	**27**	**21,355**
Nambazo Health Center	2,317	67	62	5	2,451
Kalinde Dispensary	1,946	177	16	0	2,139
Holy Family	1,966	95	61	0	2,122
Phalombe District Hospital	1,730	163	0	0	1,893
Sukasanje Health Center	1,659	71	7	0	1,737
Chitekesa Health Center	1,529	146	21	0	1,696
Migowi Health Center	1,432	6	220	0	1,658
Mkhwayi Health Center	1,217	150	127	0	1,494
Nkhulambe Health Center	1,324	50	8	4	1,386
Mpasa Health Center	1,190	111	9	2	1,312
Gogo Nazombe Health Center	1,209	86	0	1	1,296
Mwanga Health Center	687	51	1	0	739
Chiringa Maternity	501	98	3	5	607
Mulungu Alinafe	394	106	0	8	508
Chiringa Dispensary	253	28	34	2	317
Salima	14,947	1,091	466	651	17,155
Salima District Hospital	3,720	294	0	0	4,014
Khombedza Health Center	1,553	81	43	0	1,677
Chipoka Health Center	1,465	59	24	18	1,566
Lifeline Health Center	1,407	48	58	14	1,527
Thavite Health Center	1,049	69	27	93	1,238
Lifuwu Health Center	1,073	122	10	11	1,216
Maganga Health Center	746	67	12	15	840
Mchoka Health Center	704	60	31	45	840
Senga Bay Baptist Dispensary	715	58	10	0	783
Makiyoni Health Center	458	23	64	124	669
Ngodzi Health Center	364	49	17	149	579
Mafco Health Center	579	19	28	7	633
Chinguluwe Health Center	336	31	24	28	419

	VL	**EID**	**TB**	**Other**	**Total**
Katawa Health Center	282	32	60	25	399
Chitala Health Center	190	20	29	0	239
Parachute Health Centre	92	5	0	93	190
Chagunda Health Center	113	18	21	18	170
Kaphatenga Health Center	101	36	8	11	156
Rumphi	**6,657**	**483**	**1,824**	**1,378**	**10,342**
Rumphi District Hospital	1,804	66	341	5	2,216
Bolero Health Center	1,847	125	155	29	2,156
Lura Health Center	506	19	83	319	927
Mhuju Hospital	533	52	51	49	685
Katowo Rural Hospital	518	43	41	73	675
Jalawe Health Center	3	4	257	244	508
Chitsimuka Health Center	0	0	465	3	468
Nthenje Dispensary	156	10	55	176	397
DGM Livingstonia Hospital	212	22	66	67	367
Mwazisi Health Center	244	29	57	8	338
Chitimba Health Center	226	26	0	62	314
Ngonga Health Center	128	15	78	68	289
Mzokoto Health Center	238	38	1	1	278
Luwuchi Health Center	121	14	3	92	230
Mlowe Health Center	41	5	67	45	158
Mphopha Health Center	35	6	62	42	145
Tcharo Dispensary	1	0	5	94	100
Eva Demaya	44	9	37	1	91
Grand Total	**40,958**	**2,979**	**2,859**	**2,056**	**48,852**

Appendix B

Sample Transport Optimization

B.1 Optional Scheduling Constraints

For best performance, it is advisable to allow as much flexibility as possible in the construction of OST routes and schedules. However, for practical reasons it is sometimes necessary to incorporate additional constraints in response to developments in the field (e.g., urgent transportation requests) or competing objectives (e.g., fairness). Based on practical experience, the following optional constraints were incorporated into the OST model:

- Minimum weekly visit frequency at each facility:

$$\sum_{t \in \mathcal{W}_1} y_i(t) \geq \texttt{weekly\_visit\_freq}_i, \qquad i = 2, \ldots, n+1. \tag{B.1}$$

- Deadlines for the next visit to a facility:

$$\sum_{t=1}^{\texttt{deadline}_i} y_i(t) \geq 1, \qquad i = 2, \ldots, n+1. \tag{B.2}$$

- Maximum number of days between visits to a facility:

$$\sum_{\tau = t - \texttt{max\_visit\_gap}_i}^{t} y_i(\tau) \geq 1, \qquad i = 2, \ldots, n+1, \; t = 1, \ldots, T. \tag{B.3}$$

- Avoid visiting the same location on consecutive days:

$$y_i(t-1) + y_i(t) \leq 1, \qquad i = 2, \dots, n+1, \; t = 1, \dots, T. \tag{B.4}$$

B.2 Derivation of MDP Lower Bound

We approximate the infinite-dimensional MDP problem discribed in Section 3.4 as a finite-dimensional MDP with the following modifications:

- We restrict the total number of samples that may accumulate at any stage the the diagnostic network to a finite upper bound: $s_{i,k}(t) \leq s_{i,k}^{max}$. We assume that any samples/results that exceed these maximum queue lengths are discarded, i.e.

  $$s_{i,k}(t+1) = \min(s_{i,k}^{max}, s_{i,k}(t) + f_{i,k}(t) - f_{i,k+1}(t)).$$

  To compensate for the truncation of the state space, we apply a penalty to the cost function which is equivalent to the minimum number of days that the discarded samples/results would have remained in the ST system, assuming that daily transportation is available between all locations. For example, results discarded in stage 7 have a penalty of 1 day, while results discarded in stage 6 have a penalty of $1 + \omega_6$ days.

- The distance budget and all route distances are restricted to integer values. In order to obtain a valid lower bound for the infinite-dimensional problem, the integer route distances should not exceed the actual distances, and the integer distance budget should be no smaller than the actual distance budget.

Proof of MDP lower bound Let P1 be the infinite-dimensional MDP problem defined in Section 3.4:

$$P1 : \min_{\pi} \quad \lim_{T \to \infty} \frac{1}{T} \mathrm{E}[\sum_{t=0}^{T-1} R(\Gamma(t), \pi(\Gamma(t)))] \tag{B.5}$$

$$s.t. \quad S(t+1) = Q(S(t), \pi(\Gamma(t)), \mathbf{u}(t)) \tag{B.6}$$

$$b(t+1) = b(t) + \bar{d} - d(t) \tag{B.7}$$

$$w(t+1) = \mod (w(t), 7) + 1 \tag{B.8}$$

$$\pi : \mathcal{S} \to \mathcal{A} \tag{B.9}$$

Let P0 be the truncated MDP problem:

$$P0 : \min_{\pi} \quad \lim_{T\to\infty} \frac{1}{T} \mathrm{E}[\sum_{t=0}^{T-1} \hat{R}(\hat{\Gamma}(t), \pi(\hat{\Gamma}(t)))] \tag{B.10}$$

$$s.t. \quad \hat{S}(t+1) = \hat{Q}(\hat{S}(t), \pi(\hat{\Gamma}(t)), \mathbf{u}(t)) \tag{B.11}$$

$$b(t+1) = b(t) + \bar{d} - d(t) \tag{B.12}$$

$$w(t+1) = \mod (w(t), 7) + 1 \tag{B.13}$$

$$\pi : \hat{\mathcal{S}} \to \mathcal{A} \tag{B.14}$$

In problem P0, the system dynamics $\hat{Q}$ are similar to Q in P1, except that sample volumes at each location are truncated at the maximum volumes $s_{i,k}^{max}$. The cost function $\hat{R}$ is similar to R, but contains a penalty for each discarded sample:

$$\hat{R}(\hat{\Gamma}(t), \pi(\hat{\Gamma}(t))) = R(\hat{\Gamma}(t), \pi(\hat{\Gamma}(t))) + \sum_{i=1}^{n} \sum_{k=1}^{7} \rho_{i,k} \hat{z}_{i,k} \tag{B.15}$$

$$\textbf{where } \hat{z}_{i,k} = \max(0, s_{i,k}(t) + f_{i,k}(t) - f_{i,k+1}(t) - s_{i,k}^{max}). \tag{B.16}$$

In the truncated problem, the matrix $\hat{Z} = Q(\hat{S}(t), \pi(\hat{\Gamma}(t)), \mathbf{u}(t)) - \hat{Q}(\hat{S}(t), \pi(\hat{\Gamma}(t)), \mathbf{u}(t))$ represents the samples discarded at each stage in the network on day t, and the penalties $\rho_{i,k}$ are equivalent to the minimum amount of time those samples would have remained in the system (assuming that daily transportation is available between all locations). Note that $\hat{Z}$ is not part of the state space, as the discarded samples are not carried forward on subsequent days.

We now show that $opt(P0) \leq opt(P1)$, where $opt()$ represents the optimal objective value in each problem.

Let H be a hybrid problem that tracks both the truncated and infinite-dimensional versions of the ST system in parallel. Samples in the truncated system are represented by the matrix $\hat{S}$ of state variables $\hat{s}_{i,k}$, which correspond to the truncated sample volumes in P0 and follow the system dynamics described by $\hat{Q}$. All excess samples that are discarded from the truncated queues are tracked in the second system represented by state variables $z_{i,k}$. The total number of samples at each point in the ST system is $s_{i,k} = \hat{s}_{i,k} + z_{i,k}$, which corresponds to the state space of problem P1. The state space for the hybrid problem is therefore $\bar{\mathcal{S}} = \hat{\mathcal{S}} \times \mathcal{Z}$, where $\mathcal{Z} = \{Z | z_{i,k} \geq 0\}$ and $\bar{\Gamma} = (\hat{S}, Z, b, w)$.

The reward function for the hybrid problem is $\bar{R}(\bar{\Gamma}, \pi(\bar{\Gamma})) = \sum_i \sum_{k=1}^{7} \hat{s}_{i,k} + z_{i,k}$ (i.e., the cost for the truncated problem as well as the truncated samples).

$$H : \min_{\pi} \quad \lim_{T \to \infty} \frac{1}{T} \mathrm{E}[\sum_{t=0}^{T-1} \bar{R}(\bar{\Gamma}(t), \pi(\bar{\Gamma}(t)))] \tag{B.17}$$

$$s.t. \quad \hat{S}(t+1) = \hat{Q}(\hat{S}(t), \pi(\bar{\Gamma}(t)), \mathbf{u}(t)) \tag{B.18}$$

$$Z(t+1) = \bar{Q}(\hat{S}(t), Z(t), \pi(\bar{\Gamma}(t)), \mathbf{u}(t)) \tag{B.19}$$

$$b(t+1) = b(t) + \bar{d} - d(t) \tag{B.20}$$

$$w(t+1) = \mod (w(t), 7) + 1 \tag{B.21}$$

$$\pi : \mathcal{S} \hat{\times} \mathcal{Z} \to \mathcal{A} \tag{B.22}$$

We now define two versions of the hybrid problem, H0 and H1, that differ in the system dynamics function $\bar{Q}$. The dynamics of the first problem, $H0$, correspond to to the truncation penalties in problem $P0$:

$$H_0 : Z(t+1) = \quad = Q(Z(t), \mathbf{e}) + Q(\hat{S}(t), \mathbf{u}(t), \mathbf{y}(t)) - \hat{Q}(\hat{S}(t), \mathbf{u}(t), \mathbf{y}(t)), \tag{B.23}$$

where $\mathbf{e}$ corresponds to an $n+1$ vector of ones (i.e., a (possibly infeasible) route that visits all locations in the ST network in a single day). This implies that any truncated samples proceed through the ST network in a deterministic manner, and that the number of days that they remain in the network is equivalent to the penalty imposed in P0.

Second, we consider a version of the hybrid problem that corresponds to the infinite dimensional problem:

$$H1 : Z(t+1) = Q(Z(t), \mathbf{y}(t)) + Q(\hat{S}(t), \mathbf{u}(t), \mathbf{y}(t)) - \hat{Q}(\hat{S}(t), \mathbf{u}(t), \mathbf{y}(t)). \tag{B.24}$$

In this version, truncated samples move through the network according to the same ST routes as the untruncated samples.

Based on the system dynamics of problems $H1$ and $H0$, it is clear that they share the same set of feasible policies, and that the average cost of every feasible policy in problem H1 will be at least as high as the average cost of the same policy in H0 (i.e., $opt(H0) \leq opt(H1)$). We will now show that the optimal costs in these two problems are equivalent to the optimal costs in problems P0 and P1, respectively.

1) $opt(P1) = opt(H1)$

1.1) Every optimal solution in P1 can be mapped to a corresponding feasible solution in H1 with the same objective value.

Let π^* be an optimal solution in problem P1. Then, define the following policy in H1:

$$\bar{\pi}(\hat{S}, Z, b, w) = \pi^*(\hat{S} + Z).$$

The policy $\bar{\pi}$ is feasible for H1 and has the same objective value as P1. Therefore, $opt(H1) \leq opt(P1)$.

1.2) Every optimal solution in H1 can be mapped to a corresponding feasible solution in P1 with the same objective value.

Let $\bar{\pi}^*(\hat{S}, Z, b, w)$ be an optimal solution in problem H1, and define a mapping between the state space of problem P1 and H1:

$$B(S) = \{(\hat{S}, Z, b, w) | \hat{S} + Z = S\}.$$

Similarly, let

$$C(S) = \{\bar{\pi}^*(\hat{S}, Z, b, w) | \hat{S} + Z = S\}$$

and define the following feasible policy in P1:

$$\pi^*(S) = \text{ mode}(C(\mathbf{s})).$$

- If the cardinality of $C(S)$ is equal to 1 for all § in P1, then this policy has the same objective value in P1 as in H1, and therefore $opt(H1) \geq opt(P1)$.

- If the cardinality of $C(S)$ is greater than 1 for any S in P1, then there exists a pair of states for which $(\hat{S}^{(1)}, Z^{(1)}) \in C(S)$ and $(\hat{S}^{(2)}, Z^{(2)}) \in C(S)$ such that

$$\bar{\pi}^*(\hat{S}^{(1)}, Z^{(1)}) = \mathbf{y}^{(1)} \neq \mathbf{y}^{(2)} = \bar{\pi}^*(\hat{S}^{(2)}, Z^{(2)}).$$

  Since the cost function in problem H1 depends only on the total number of samples in the system (i.e., $\hat{S}^+ Z$ and $\bar{\pi}^*$ is an optimal policy, it follows that actions $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ must have identical value functions for both states. We can therefore assign either of these actions to both states without changing the average cost of the policy. We repeat this process until $|C(\hat{S})| = 1$ for all $\hat{S}$ to obtain an equivalent solution for problem P1 with the same objective value. Therefore $opt(H1) \geq opt(P1)$.

Combining the inequalities above, we conclude that $opt(P1) = opt(H1)$.

2) $opt(P0) = opt(H0)$

2.1) Every optimal solution in P0 can be mapped to a corresponding feasible solution in H0 with the same objective value.

Let $\hat{\pi}^*$ be an optimal solution in problem P0. Then, define the following policy in H0:

$$\bar{\pi}(\hat{S}, b, w) = \hat{\pi}^*(S, 0 * Z, b, w).$$

The policy $\bar{\pi}$ is feasible for P0 and has the same objective value as H0. Therefore, $opt(H0) \leq opt(P0)$.

2.2) Every optimal solution in H0 can be mapped to a corresponding feasible solution in P0 with the same objective value.

Let $\bar{\pi}^*(\hat{S}, Z, b, w)$ be an optimal solution in problem H0, and define a mapping between the state space of problem P0 and H0:

$$B(\hat{S}) = \{(\hat{S}, Z) | Z \in \mathcal{Z}\}.$$

Similarly, let

$$C(\hat{S}) = \{\bar{\pi}^*(\hat{S}, Z) | Z \in \mathcal{Z}\}$$

and define the following feasible policy in P0:

$$\pi^*(\hat{S}) = \text{ mode}(C(\hat{S})).$$

- If the cardinality of $C(\hat{S})$ is equal to 1 for all $\hat{S}$ in P0, then this policy has the same objective value in P0 as in H0, and therefore $opt(H0) \geq opt(P0)$.
- If the cardinality of $C(\hat{S})$ is greater than 1 for any $\hat{S}$ in P0, then there exists a pair of states $(\hat{S}^{(1)}, Z^{(1)}) \in C(\hat{S})$ and $(\bar{S}^{(2)}, Z^{(2)}) \in C(\hat{S})$ such that

  $$\hat{\pi}^*(\hat{S}^{(1)}, Z^{(1)}) = \mathbf{y}^{(1)} \neq \mathbf{y}^{(2)} = \hat{\pi}^*(\hat{S}^{(2)}, Z^{(2)}).$$

  Since the cost function in problem H1 depends only on the number of samples in the truncated system (i.e., $\hat{S}$ and $\bar{\pi}^*$ is an optimal policy, it follows that actions $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ must have identical value functions for both states.
  We can therefore assign either of these actions to both states without changing the average cost of the policy. We repeat this process until $|C(\hat{S})| = 1$ for all $\hat{S}$ to obtain an equivalent solution for problem P1 with the same objective value. Therefore $opt(H0) \geq opt(P0)$.

Combining the inequalities above, we conclude that $opt(P0) = opt(H0)$.

B.3 Details on Operational Implementation

In this section, we provide more operational information about the field implementation. First, we describe the personnel involved with the implementation and their responsibilities (Appendix B.3.1). We then describe the three main data sources that provide input data for the OST system (Appendix B.3.2).

B.3.1 Field Staff

The implementation of the OST system was overseen by a local field manager based at R4H's central office in Lilongwe, as well as three research assistants based in the implementation districts. Research assistants were responsible for monitoring and addressing issues associated with daily data collection from both couriers and healthcare facility staff in their district. The field manager served as a central point of contact between the international research team and local staff involved in the implementation, including R4H couriers and regional transportation coordinators, as well as district officials, diagnostics staff, and laboratory managers.

B.3.2 Data

As described in Section 3.4, the the ST network model requires two sets of input parameters (the vectors of arrivals and initial location of samples; $\mathbf{u}$ and $\mathbf{v}$) to formulate the daily route optimization problem. During January – July 2019, we worked closely with R4H staff to identify relevant existing data sources within the diagnostic network and implement new data collection platforms to obtain the required input data.

B.3.2.1 Forecasting Sample Volumes

To estimate $\mathbf{u}$, the future sample volumes at each healthcare facility, we developed a simple forecasting model based on average historical sample volumes on each day of the week. This strategy was particularly effective for viral load samples, which account for over 90% of the samples transported by R4H couriers. Viral load monitoring was generally conducted on 1–3 designated days of the week at each healthcare facility, resulting in easily identifiable patterns in incoming sample volumes.

B.3.2.2 Monitoring Current Sample Volumes

In order to monitor the current sample volumes at each location within the ST system, we relied on a combination of three data sources within the diagnostic network, including the USSD application described in Chapter 2. We briefly describe each of these data sources below.

USSD Application. Sample volumes at healthcare facilities were monitored via the USSD application described in Chapter 2. In conjunction with laboratory managers in each district, we recruited healthcare facility staff to participate in USSD sample volume reporting and facilitated initial training sessions at each district hub. During these sessions, R4H staff presented an overview of the OST system and explained the need to monitor sample volumes at each facility. Staff were then guided through a hands-on demonstration of the USSD application and provided with informational posters to display at their facilities. All staff present were able to access the system via their personal phones, and most found the reporting process simple and easy.

Following the initial training sessions, district research assistants conducted in-person visits at healthcare facilities to assess reasons for irregular participation and take corrective actions. For example, additional staff were trained to use the USSD application at facilities where the original participants were no longer directly involved in sample collection.

Sample Tracking Application. Once samples were collected by a R4H courier, it was significantly easier to track their continued progress through the diagnostic network. ST couriers maintained detailed internal logs to record when each sample and result was moved between different locations within the ST system. In order to facilitate real-time monitoring of this data, we worked closely with R4H to design a comprehensive tablet-based sample tracking application using the CommCare mobile platform. Starting from April 2019, ST couriers used this application to submit daily sample transportation logs via Android tablets, and the corresponding data was stored in R4H's centralized ST database.

Laboratory Data. Processing of samples and results at molecular laboratories was tracked via the national Laboratory Information Management System (LIMS). Laboratory logs were used by the OST model to monitor the number of results available for collection at laboratories, and estimate the number of results to be printed in subsequent days based on current testing throughput.

B.3.2.3 Data Reliability

Each of the three data platforms used to monitor ST operations experienced occasional outages during the implementation period due to external factors such as mobile network failures. To mitigate the effect of these disruptions, we developed simulation-based heuristics and imputation methods to estimate missing data required for daily route optimization. We also implemented a number of automated data cleaning procedures to monitor for inaccurate or inconsistent data and remove errors from the OST model input.

B.4 ST Delays Counterfactual Estimation Procedure

In order to contrast the efficiency of OST routes with fixed routes, counterfactual fixed schedule ST delays were calculated for samples and results transported during the OST pilot study.

Counterfactual delays were estimated using the fixed weekly ST schedules that were in place prior to the OST pilot study, translated to the period during which the study took place (and accounting for public holidays, courier leave days, strikes/protests, etc. in the study period).

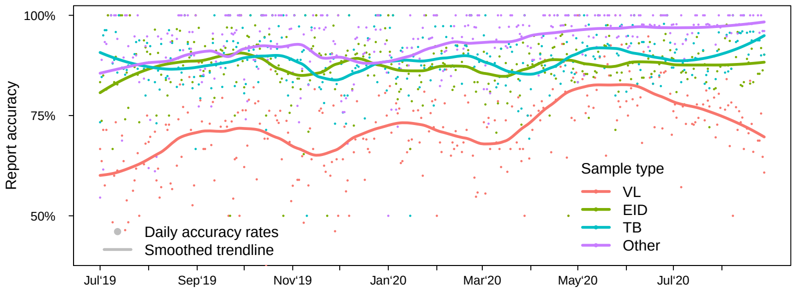
From actual sample data logged during the pilot study, we extracted the date on which each sample was created at the healthcare facility. We assume that samples delivered to a district hub are ready for delivery to the molecular laboratory the morning after they arrive. From this and the fixed weekly ST schedule, we calculate a counterfactual delay for the sample to reach the molecular laboratory from its creation at the health facility.

Likewise, from actual results data logged during the pilot study, we extract the date on which each result was created at the molecular laboratory. Again assuming that results delivered to a district hub are ready for delivery to health facilities the next morning, we used the fixed weekly schedule to calculate a counterfactual delay for a sample to reach its health facility from the molecular laboratory.

B.5 Additional Results for USSD Data Accuracy

Approximately 85% of sample volume reports submitted via the USSD application were consistent with sample tracking records maintained by R4H couriers. The proportion of inconsistent reports decreased over time, ranging from 19% in July-August 2019 to 13% during the same period in 2020. As illustrated in Figure B.1, reporting accuracy was lowest for viral load samples and highest for "Other" samples. The discrepancy in accuracy across different sample types was primarily due to the higher sample volumes for viral load testing. Facility staff were approximately six times more likely to report incorrectly when there was at least one sample present at the facility, relative to days when there were no samples at the facility. The most frequent explanations for incorrect reports were as follows: (1) facility staff forgot to count older samples that had accumulated on previous days; (2) samples were removed from the facility or additional samples were collected from patients after the report was submitted; (3) facilities reported items that were not samples (e.g., monthly reports were counted in "Other" samples).

Figure B.1: Accuracy of USSD reports, July 2019 – August 2020.



B.6 Additional Results for Unnecessary Visits

Figure B.2 summarizes the most common factors that contributed to unnecessary trips in the OST schedules. On average, missing or incomplete USSD data contributed to just over half of all unnecessary visits: 30% of unnecessary trips were associated with incorrect USSD reports in the preceding days (i.e., the facility reported new samples, but none were found by the courier), and 45% of unnecessary trips were associated with incomplete USSD reports (the facility did not report whether they had any samples). Approximately 29% of unnecessary trips took place at least a week after the last visit to that facility, indicating that minimum visit frequency constraints may have contributed to the decision to visit that location.

Figure B.3 demonstrates that unnecessary trips occurred more frequently at facilities with lower sample volumes, leading to a significantly higher proportion of unnecessary visits in Rumphi.

Finally, as illustrated in Figure B.4, the proportion of unnecessary trips decreased over time as the minimum visit frequency targets were relaxed. The model initially attempted to visit each facility once every seven days in 2019, and this requirement was gradually relaxed to one visit per calendar week at the start of 2020, and then removed altogether in mid-2020. These changes, together with improvements in USSD data quality, reduced the average proportion of unnecessary trips to approximately 2% in July – August 2020.

Figure B.4 also highlights the relatively high proportion of unnecessary trips in December 2019 and April – May 2020. During these periods, diagnostic services at healthcare facilities were significantly disrupted due to seasonal holidays and the onset of the COVID-19 pandemic, respectively. The increased frequency of unnecessary trips in these months is linked to the sudden drop in demand for sample transportation due to significantly lower sample volumes.

Figure B.2: Reasons for unnecessary visits in OST routes

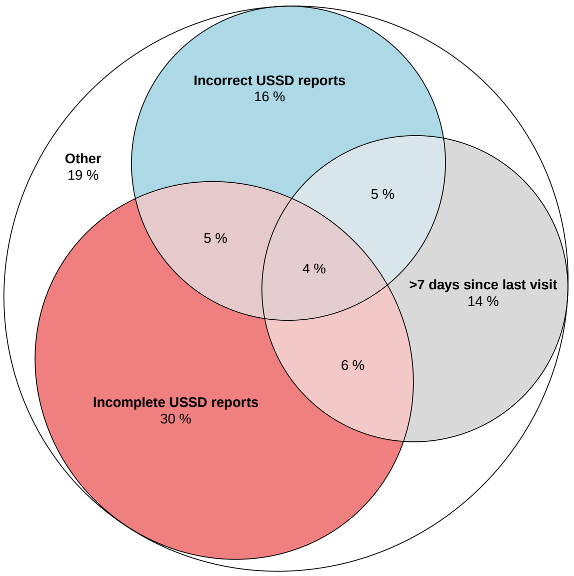


Figure B.3: Sample volumes and unnecessary trips by facility

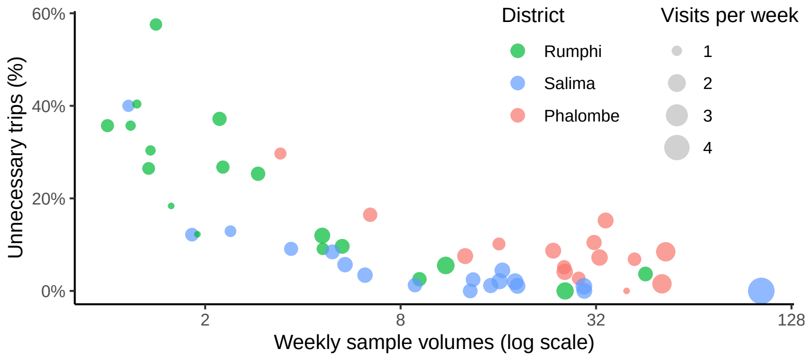
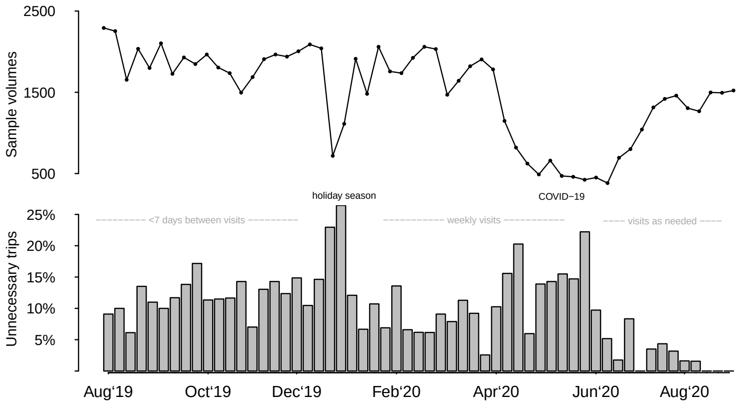


Figure B.4: Sample volumes and unnecessary trips, August 2019 – August 2020



Appendix C

Point-of-Care Technology for VL Monitoring

C.1 POC Cost Model

C.1.1 POC Testing Equipment

Each POC testing device included in our model is represented by a device index $j = 1, \ldots, m$. For each POC device we supply the following input data:

- Daily test capacity—assuming that testing is performed for approximately 7 hours per day, 250 days per year.
- Expected lifespan of the device (years)
- Fixed costs:
  - Instrument and accessories cost (per device)—this includes testing modules as well as additional equipment such as laptops, tablets, printers, centrifuges, batteries, solar panels, etc.
  - Warranty and/or Maintenance cost (per instrument, per year).
  - Delivery, installation, and calibration cost (per instrument, once-off)
- Variable costs (per test):

- Test cartridges / reagents
- Sample collection kits, PPE, and other consumable supplies
- Electricity, water, and other utilities
- Staff time, including sample collection and analysis, and result communication and documentation.

We summarize the costs associated with each instrument into total fixed and total variable costs, with fixed costs incurred regardless of device utilization, whereas variable costs are incurred on a per-test basis.

C.1.1.1 POC Data Sources

We identified four POC testing instruments currently available for purchase and suitable for VL monitoring in low-resource settings. Three instruments were from Cephid's GeneXpert range [44], which includes a variety of modular equipment configurations capable of testing 4–80 samples simultaneously, with a runtime of approximately 90 minutes per test. We also considered the Abbott m-PIMA machine [2], which runs a single test in approximately 60 minutes. We excluded less common POC testing instruments such as the SAMBA devices [61] due to limited data on their use in clinical settings.

To estimate the POC instrument costs we compared data reported in a variety of different sources, including manufacturer websites [44, 2], online procurement databases [71], and published reports [140], and academic publications [142, 37, 203]. Our fixed cost estimates are shown in Table C.1, and variable cost estimates are shown in Table C.2.

Where possible, these estimates were based on real-world implementation costs reported during POC trials such as [203], which describes costs associated with implementing EID POC testing in Zambia. We also relied on estimated costs in other modeling studies such as [142], which provides estimated costs for POC EID testing in Zimbabwe, and [37], which estimates costs for POC VL testing in Kenya.

Table C.1: POC instrument costs

	GeneXpert		GeneXpert	GeneXpert			Abbott
	II	II	IV	XVI	XVI	XVI	m-PIMA
Modules	1	2	4	8	12	16	1
Instrument	9,420	[a]12,530	[a]19,000	[a]43,000	[a]58,44	[a]71,850	[b]25,000
Software	[b]5,500	5,500	5,500	5,500	5,500	5,500	0
UPS/battery	6,000	6,000	6,000	10,000	14,000	18,000	0
Warranty/ Maintenance	[a]4,500	4,500	[a]6,840	[a]18,504	18,504	18,504	[b]9,000
Delivery/ installation	3,500	3,500	3,500	4,500	4,500	4,500	[b]4,4200
Training	2,000	2,000	2,000	2,000	2,000	2,000	2,000
Lifespan	5	5	5	5	5	5	5
Daily Capacity	4	8	16	32	48	64	6

[a] [71] [b] [142]

Table C.2: POC variable costs

	GeneXpert	Abbott m-PIMA
Cartridge & consumables	14.90	25.00
Utilities (electricity & water)	0.10	0.10
Sample kit	2.00	2.00
Staff time	2.26	2.00
Total	$19.26	$29.10

C.1.1.2 Notes on Fixed Costs

Equipment Costs. Based on recent literature, it appears that the total fixed costs associated with POC testing equipment are generally 2–3 times higher than the cost of the instrument itself. For example, the most common POC instrument is the 4-module GeneXpert device which is mentioned in [203], [142] and [37]. This instrument costs approximately $17 500, but all three studies estimate significant additional costs such as shipping, installation, software, maintenance, accessories, and power supply. Including all of these additional expenses, the total cost associated with the device ranges from $34,000 – $38,000 in these papers.

Power Supply. For 4-module GeneXpert devices, we assumed a cost of $6,000 for a battery and UPS. [142] included a battery cost of $5,500, while [37] estimated $4,000 for a battery and $150 for UPS, and [203] recorded $790 for UPS. For larger GeneXpert devices, we assumed that power supply costs would be approximately $2,000 + $1,000 per test module. We did not include power supply costs for the m-PIMA machine, which has a built-in battery [142].

Shipping and Installation. [142] estimated $1,522 for freight (insurance and customs clearance) plus $1,925 for storage and distribution for a GeneXpert IV instrument and battery, while [203] estimate $1,350 for freight and $1,800 for installation and training for the same device. We estimated similar costs of $3,500 for the 2- and 4-module GeneXpert instruments and a slightly higher cost of $4,500 for the larger 16-module machine to account for higher shipping weight, insurance, and customs duties.

Training. There was significant variability in training and infrastructure costs associated with POC devices in different studies—[142] estimated approximately $500 for training, $500 for monitoring and supervision over 5 years, while [37] included significantly higher training costs of $4,800, and [203] included a combined training and installation cost of $1,800. We assume a fixed cost of $2,000 per facility regardless of the type of instrument.

Facility Infrastructure. [203] estimated a cost of $2,043 for facility infrastructure, [142] estimated $404 for facility upgrades. In practice, the cost of facility infrastructure upgrades to support POC testing is likely to vary significantly at different locations. For example, most larger facilities (especially district hospitals) are likely to have some existing laboratory infrastructure, whereas smaller clinics that do not perform any diagnostic testing may require significant upgrades. We ultimately decided to omit this cost from our calculations, as upgrades to facility infrastructure have implications beyond the VL monitoring program and would likely need to be planned and evaluated in a much broader context.

C.1.1.3 Notes on Variable Costs

Staff and Labor Costs. Most existing literature models labor costs on a per-test basis, assuming that staff members responsible for POC testing will spend only a fraction of their work day on these activities. This assumption may be problematic in small facilities where staff do not have the appropriate qualifications to operate the equipment, requiring a new staff member to be assigned to the facility.

[142] reported an average of 27 minutes labor per POC EID sample, most of which was associated with patient interactions (drawing sample, communicating results) or administrative tasks (documentation and labeling). Only 5 minutes was allocated to POC-specific activities, which were performed by either a nurse ($4.3-$4.9 per hour) (Abbott machine) or a lab technician ($6.35 per hour) (GeneXpert). [37] estimated 8 minutes of hands-on work by a lab technician for each GeneXpert POC test, at a cost of $4.8 per hour. In both cases, there is a similar labor cost of $0.6–$0.7 for the sample testing. [151] estimated costs of $0.37 for DBS and $0.87 for plasma samples, including sample collection and testing. [14] estimated approximately 5 minutes for sample collection and referenced hourly wages of $1.09 for nurse assistants and $2.40 for laboratory technicians in Malawi, which are substantially lower than estimates in other countries and likely include only a portion of the total labor costs (e.g., benefits,

overhead, PTO, etc.).

We assumed the following labor for each POC sample:

Sample collection and documentation: 10 minutes; Transport to/from laboratory: 2 minutes; Testing: 8 minutes; Communication and documentation of results: 10 minutes. We estimated a total labor cost of $2 for m-PIMA samples (30 minutes at $4/hour) and $2.26 for GeneXpert samples (22 minutes at $4/hour, plus 8 minutes at $6/hour).

Sample Collection Kits and Consumables. [142] estimated $30.55 for consumables related to m-PIMA samples. A substantial portion of this cost is the test cartridge ($25), freight ($1.74) and storage/distribution ($2.86). For GeneXpert samples, cartridge costs are $14.9 with lower freight ($1.06) and storage/distribution ($1.79). In both cases, sample collection supplies accounted for only $0.97 of the total variable costs. [37] estimated a cost of $0.62 for sample collection kits (GeneXpert), while [48] estimated a cost of $1.19 for supplies for collecting samples to be processed on the Alere machine (similar to m-PIMA), and $2.17 for GeneXpert samples. [14] estimated $2.60 for a DBS sample collection kits in Malawi.

C.1.1.4 Combining Multiple POC Device Options

In addition to the 7 instrument configurations described in Table C.1, we also model combinations of multiple GeneXpert instruments which can be used to increase the total POC capacity. For example, the largest healthcare facilities may require 2-3 POC instruments to supply enough capacity to test a large proportion of their VL sample volumes. We do not allow combinations of GeneXpert devices with m-PIMA devices, as the m-PIMA devices have very low capacity and are unlikely to be cost-effective at any facility with high enough volumes to require multiple devices.

In cases where multiple devices are allocated to the same site, we assume that most fixed costs will be cumulative (i.e., full price will be paid for each instrument and its warranty/accessories). However, the total training cost will remain at $2,000 regardless of the number of additional devices and the delivery and installation costs for each additional instrument will be discounted by $1,000.

C.1.1.5 Centralized Testing Costs

Like POC testing, centralized testing costs usually include fixed costs for laboratory equipment, maintenance, and overheads, as well as variable costs for sample collection kits and consumables, test reagents, and staff time. Sample transportation costs are generally included as an additional variable cost (i.e., per sample), although it is likely that a substantial portion of the transportation costs are fixed based on the distances traveled and frequency of travel to each healthcare facility.

In our analysis, we estimated centralized testing costs by following a similar procedure to the POC cost estimates. We assumed that centralized testing would be performed on Abbott m2000 instruments, which cost approximately $170,000 with estimated annual maintenance costs of $18,000 and a 7-year lifespan [14]. Similar to the POC device estimates, we assumed that the total fixed costs would be at least double the cost of the instrument over its expected lifespan, i.e., approximately $60,000 in additional costs for shipping and installation, training, and laboratory infrastructure (e.g., generators, printers, computers and servers, connectivity).

Given that centralized laboratories process large volumes of samples, we assumed that the centralized instruments would operate at full capacity (93 samples per day, 250 days per year), resulting in an average fixed cost of $2.18 per sample.

Variable costs per sample include approximately $13 for test reagents and consumables ([14]), as well as $2 for sample collection kits and $0.1 for other utilities (assumed to be the same as POC samples). For labor costs, we assumed that centralized samples would require a total of 30 minutes ($2) of staff time at the facility to account for sample collection, documentation, and packaging for transportation, as well as sorting and documentation of results on delivery, and communication of results to patients. We assumed additional labor costs of $1 by district and central laboratory staff, including sorting and packaging samples and results, capturing sample and patient details in the laboratory database, sample viability checks and quality control, and testing. Finally, we assumed transportation costs of $0.5 per item per trip. This includes vehicle-related costs (fuel, maintenance, etc.), packaging to protect samples and results during transportation, and courier labor (travel time as well as sorting and documentation of items transported). Most healthcare facilities had total transportation costs of $2 per test for transportation of both the sample and result between the facility, district hub, and laboratory. Samples originating at facilities with molecular laboratories did not incur any transportation charges, while those originating at district transportation hubs or at facilities in districts with molecular laboratories incurred a transportation cost of $1 per test (for comparison, [14] considered a range of $1.49–$2.24).

The total estimated cost per centralized test is therefore $20.28–$22.28, depending on the amount of transportation required. Based on VL sample volumes in 2021, the estimated average cost was $21.56.

Other cost estimates for centralized VL monitoring vary significantly in the recent literature. [173] assume a cost of $22 per centralized test (Zimbabwe, 2017), including $3 for staff time, $2 for collection kits and consumables, $2 for communicating test results, and $15 for laboratory costs (reagents, equipment, maintenance, etc). [37] assume a centralized test cost of $24.63 based on 2017 USAID estimates for Kenya, which include sample transportation. [151] estimated costs of $18 – $23 per test (Zambia, 2019), including $0.31 – $5 for sample collection kits (with PSC kits being substantially more expensive than Dry Blood Spots), $0.47-$0.87 for sample collection (staff time, equipment, and overhead) and $17.22-$17.54 for laboratory analysis. [14] estimated an average cost of $19.39 ($17.73) for centralized testing using DBS (or PSC) (Malawi, 2018), including $2.6 ($4.5) for sample kits, $13 ($9.4) for reagents and test consumables, $1.87 for sample transportation, as well as fixed costs of $170,000 ($300,000) for testing equipment processing approximately 47 (18) samples per hour, with annual maintenance costs of $18,000 ($30,000) over a 7-year lifespan. [80] listed costs of $24.46 for centralized testing and $24–$38 for the GeneXpert and m-PIMA machines, based on 2017 South African National Laboratory pricing. [77] quoted an international benchmark for centralized test costs of $28.62.

C.1.1.6 Near-POC Costs

We estimated near-POC costs based on a combination of the POC and centralized testing costs. Near-POC equipment costs were modeled based on the instrument costs in Table C.1, with an additional cost of $200 per year for each facility referring samples for near-POC testing (to cover administrative costs and training). For simplicity, we assumed that only GeneXpert devices would be used for near-POC testing, and that near-POC samples would incur the same variable costs for cartridges, utilities, and sample collection kits. We assumed $2 of staff labor at the healthcare facility, $0.5 at the testing site, and $1 for transportation to and from the healthcare facility.

C.1.2 Healthcare Facilities in Malawi

Healthcare facilities are denoted using the index $i = 1, \ldots, n$. For each healthcare facility in the model, we require the following input data:

- The district and location of the healthcare facility
- Whether the facility is a Molecular Laboratory, Central Hospital, District Hospital, or transportation hub. These characteristics influence the estimated cost of centralized testing, as well as whether the facility is eligible to test near-POC samples from other facilities.
- The approximate HIV prevalence in the communities surrounding the facility.
- The annual VL sample volumes.
- The distribution of VL sample volumes on different days of the week (generally determined by the days on which ART clinic services are available).

C.1.2.1 Data Sources

Our primary data source for healthcare facilities in Malawi was the national Laboratory Information Management System (LIMS) [126]. This database contains records of all

VL samples tested in any of the 10 national molecular laboratories, which collectively process 99.8% of VL tests conducted in Malawi [17].

Healthcare Facilities Offering VL Testing. We extracted a list of healthcare facilities in Malawi offering VL monitoring services from the LIMS records. We included all facilities that referred at least one sample for VL monitoring in 2021. The total number of facilities included in the analysis was 702. Note that this is substantially lower than the number of healthcare facilities in Malawi, as some facilities do not offer VL monitoring while others may have access to private laboratory equipment.

Facility Location Data. We determined the district of each healthcare facility based on the LIMS data and we obtained GPS coordinates for healthcare facilities from the Master Health Facility Registry [127] and other online sources [18]. In cases where exact coordinates were unavailable, we used approximate coordinates corresponding to the village, town, or nearby landmarks such as schools and churches.

HIV Prevalence Data. We used geospatial data from the 2015-2016 Malawi Demographic and Health Survey [130] to generate a heatmap of interpolated HIV prevalence rates across Malawi. This analysis was performed using the R package prevR and an approach similar to the methods described in [155]. Based on the interpolated prevalence map, we extracted estimated HIV prevalence rates at the GPS coordinates associated with each healthcare facility (see Figure C.1).

While this analysis captures general variations in HIV prevalence across different areas, the approach has several important limitations. The HIV prevalence estimates do not necessarily reflect populations in the catchment area associated with specific healthcare facilities, especially in densely populated areas where patients may have access to multiple facilities. The data also corresponds to 2016 HIV prevalence rates, which may not have changed in different ways after recent interventions (we anticipate that more recent data from the 2020-2021 DHS will be available in 2022 or 2023).

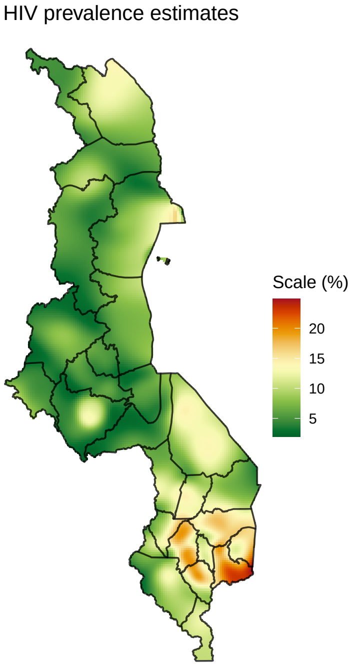


Figure C.1: Interpolated HIV prevalence rates among 15-49 year-olds in 2016.

C.1.3 Supplementary Results: POC Cost Analysis

Table C.3 summarizes the amount of POC capacity allocated to healthcare facilities for each of the scenarios considered in Section 4.3.1, while Figure C.2 illustrates the relationship between annual sample volumes and POC testing costs in scenarios where facilities perform a 100% POC testing without cost-sharing. These data indicate that expanding facility schedules to allow testing on every day of the week allows many facilities to achieve 100% coverage with smaller instruments, resulting in lower costs.

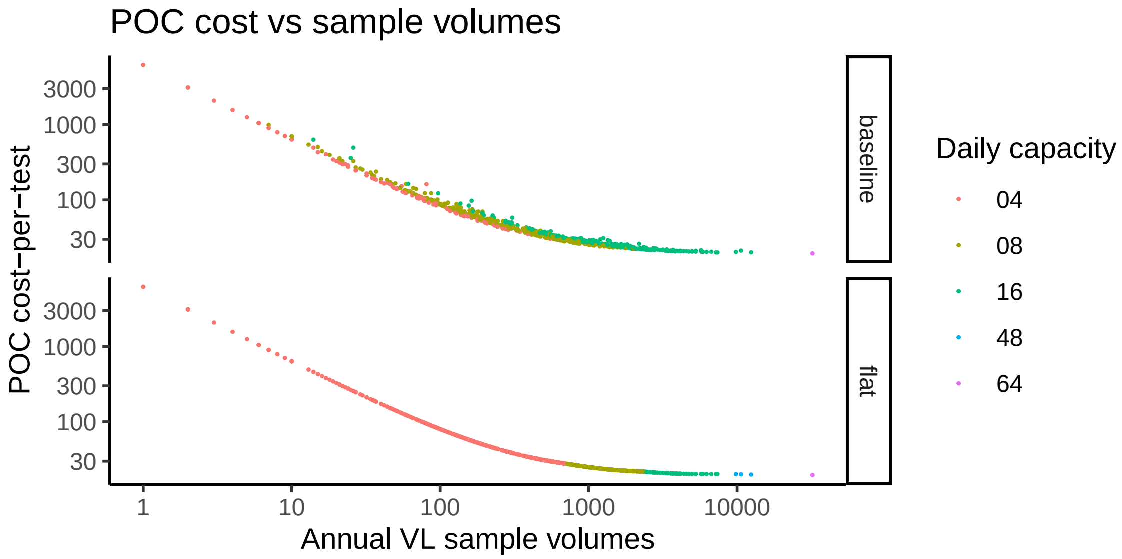


Figure C.2: A comparison of POC cost-per-test vs. annual sample volumes at healthcare facilities in Malawi. Each point represents a single facility, and the illustrated cost estimates correspond to the instrument with the lowest cost-per-test (without cost-sharing) for 100% POC testing. The capacity of the selected instrument is indicated by the color of each point. The first graph is shows costs for the baseline demand patterns in the 2021 data, while the second graph shows estimates for testing performed 5 days per week.

Table C.3: A summary of the number test modules allocated to healthcare facilities for the operational strategies in Table 4.5. Each test module is assumed to provide 4 tests per day. The GeneXpert instruments corresponding to each number of modules are listed in Table 4.1.

	Shared	Baseline schedules						Expanded schedule					
	capacity	1	2	4	8	12	16	1	2	4	8	12	16
POC only	0%	62	98	194	242	76	29	301	227	125	42	4	3
	25%	43	57	166	245	114	75	220	215	174	76	11	6
	50%	28	34	98	194	157	188	101	200	227	125	32	16
	75%	9	19	34	98	106	406	13	88	200	227	81	90
POC + centralized	0%	118	338	245	0	0	1	448	190	60	0	3	1
	25%	86	256	357	0	0	3	361	236	99	1	2	3
	50%	45	181	466	0	4	6	259	270	158	0	9	6
	75%	11	92	548	0	10	41	83	261	304	0	22	32

C.2 HIV Synthesis Simulation Model

In this section we provide additional details on the methods used to simulate the impact of introducing POC and near-POC VL monitoring at healthcare facilities in Malawi. We begin with an overview of the HIV Synthesis simulation model, followed by a description of how the model was calibrated for Malawi's population in Appendix C.2.2. In Appendix C.2.3 we discuss the parameters and assumptions used to model the POC and near-POC VL monitoring.

C.2.1 HIV Synthesis Model Overview

The HIV Synthesis model is an stochastic simulation that models the transmission, diagnosis, and treatment of HIV within a population comprising thousands of individuals. The simulation is initialized in the year 1984 with a small number of HIV infections among the initial population, and the model is then updated in 3-month increments to simulate the spread of HIV through heterosexual transmission among individuals in the population. The model also simulates the implementation of various public health policies in response to the epidemic at the appropriate points in time (e.g., HIV testing and ART treatment programs).

General descriptions of the model is and its applications can be found in previous studies such as [169, 175, 173, 174] and the associated supplementary material, and comprehensive documentation of the model can also be viewed at `hivmodeling.org/model-database/hiv-synthesis`. In the remainder of this section, we limit our discussion to the components of the model that are directly relevant to our assessment of POC testing strategies.

Time Horizon. For the analysis reported in this work, we focused on two key time periods: the pre-intervention period (2012–2022) and the intervention period (2023–2028). During the pre-intervention period, we calibrated the model assumptions and parameters to reflect the characteristics of Malawi's population and the evolution of Malawi's HIV treatment policies. During the intervention period, we simulated a variety of different operational strategies for introducing POC VL monitoring, and we used simulation output from this period to evaluate the cost-effectiveness of these strategies.

We selected an intervention period of 5 years to match the expected lifespan POC testing equipment. This enables us to draw concrete comparisons between the fixed costs associated with procuring POC instruments and the expected benefits of the instruments over their useful lifespan, without significant additional assumptions about the availability and cost of POC technology in future. One limitation of this approach is that the outcomes simulated over a period of 5 years may not capture some of the long-term benefits of improved VL monitoring (e.g., fewer new infections, resulting in decreased mortality and morbidity over a 20-30 year horizon. It is therefore reasonable to expect that the impact estimated derived from this analysis are smaller that other analysis performed with the HIV synthesis model over longer horizons (e.g., 20 years in [169], 10 years in [170]).

C.2.2 Model Calibration

The HIV Synthesis model includes hundreds of parameters that are used to simulate population demographics, sexual behaviour, HIV transmission, disease progression, and individual health outcomes, as well as the effect of HIV prevention programs, testing, and treatments. Extensive work has been done to select appropriate parameter ranges for various settings in sub-Saharan Africa, including South Africa, Zimbabwe [169, 173] and Malawi [171, 172]. The parameters used in our analysis are generally consistent with this prior work, except in cases where it was necessary to alter our approach to reflect different VL monitoring policies.

Our general approach to the model calibration was similar to methods used in previous work—we initially generated a large number of simulation instances with parameters sampled from appropriate distributions, and then selected a smaller subset of these instances for our final analysis. Since our work focuses on differentiated policies at the level of individual healthcare facilities, our sampling criteria were slightly more complex than those used in previous work (which has generally focused on national populations). In particular, the final set of simulation instances were selected to ensure that there were sufficient instances to reflect the characteristics of each healthcare facility, as well as the overall demographics of Malawi's population.

C.2.2.1 Facility Calibration

Our analysis assumed that each healthcare facility is associated with an independent population consisting of both HIV-positive and -negative individuals, and that the effects of implementing different VL monitoring strategies at the facility are limited to individuals in the corresponding population (i.e., the outcomes associated with each facility are independent of operational choices at all other facilities). We estimated the effect of different VL monitoring strategies in the facility populations by matching each facility to at least 20 simulation instances and aggregating the effects of each policy in those simulations.

HIV Prevalence. To match facilities to appropriate simulation instances, we focused primarily on the facility HIV prevalence estimates described in Appendix C.1.2.1. We clustered both the facilities and the simulation instances into 24 groups based on 1% HIV prevalence intervals (i.e., 2-3% HIV prevalence in the lowest group and 24-25% HIV prevalence in the highest group). We then matched each group of facilities to the corresponding group of simulation instances with similar HIV prevalence rates in the year 2016, and used the output from this set of simulations to estimate the impact of implementing various VL monitoring policies at the facilities.

Facility Scale Estimates. To account for differences in population size at each healthcare facility, we multiplied the simulation output by a scaling factor of $\frac{A_i}{B_i}$, where A_i is the total number of VL monitoring tests conducted for patients at the facility in 2019–2021 (from the national laboratory data), and B_i is the total number of successful VL tests over the same period in the matched simulation scenarios. This approach assumes that the number of PLHIV associated with each facility is proportional to the number of VL tests conducted at the facility, and that the total

Table C.4: Summary of national population and HIV statistics used for model calibration.

Year	Statistic	Source	Observed value	Simulated value
2018	Population 15+	National census [131]	9 845 162	9 845 102
2018	Population 15-64	National census [131]	9 188 275	9 373 908
2018	Population 15-64	World bank estimate [230]	10 400 268	9 795 338
2020	PLHIV (15+)	UNAIDS estimates [213]	930 000	946 000
2020	PLHIV (15+)	MPHIA 2020 [167]	946 000	946 000
2020	PLHIV (15-49)	IHME [97]	744 871	730 983
2019	People receiving ART (15+)	PEPFAR [160]	763 588	763 588
2020	People receiving ART (15+)	PEPFAR [160]	773 178	777 621
2021	People receiving ART (15+)	PEPFAR [160]	819 956	786 076
2019	HIV prevalence (15-49)	IHME [97]	8.4	8.5
2020	HIV prevalence (15-49)	CDC [43]	8.1	8.1
2020	HIV prevalence (15+)	MPHIA 2020 [167]	8.9	9.2
2021	HIV prevalence (15+)	PEPFAR [160]	9.1	9.0

population associated with each facility is proportional to the number of PLHIV, after accounting for different HIV prevalence rates.

C.2.2.2 National Calibration

We modeled the national population of Malawi by aggregating over the simulated populations for each facility. We calibrated the aggregated simulation statistics to reflect key population and HIV statistics from a variety of sources, which are summarized in Table C.4.

C.2.3 VL Monitoring Parameters and Assumptions

In this section we provide additional details on the parameters and assumptions used to model different types of VL monitoring in the HIV Synthesis simulations.

C.2.3.1 Scale-Up of VL Monitoring

We assumed that VL monitoring of ART patients was introduced in the years 2013, 2014, 2015, 2016, 2017, or 2018 with the following probabilities: (0.15, 0.30, 0.40, 0.075 ,0.05, 0.025). These probabilities are based on historical records from the national LIMS database, and are approximately proportional to the number of facilities that began referring VL samples in each in year.[*].

We simulated various changes to VL monitoring policies over the years 2013–2020, in line with Malawi's treatment guidelines. The initial threshold for a high VL result was 5000 copies/ml, which changed to 1000 copies/ml in 2016. Routine VL monitoring was conducted every 24 months prior to 2019, and annually thereafter.

C.2.3.2 Probability of VL Monitoring Success

Previous versions of the HIV Synthesis simulation have used a single parameter (eff_prob_vl_meas_done) to model the probability that a patient who is eligible for VL monitoring under the relevant treatment guidelines will actually receive a successful VL monitoring test. This parameter accounts for the effect of a number of different factors that can interfere with VL monitoring—the clinician may forget or decide not to collect a sample, the sample may be lost, damaged, or non-viable, or the test result may be lost or omitted from the patient's records.

In our analysis, we used a more detailed model of VL "success rates" to reflect differences in centralized, POC, and near-POC testing, as well as improvements in the provision of diagnostic services over time. We assumed that the success rate of VL monitoring tests was determined by three parameters:

$$\text{eff\_prob\_vl\_meas\_done} = p_{sample} \times p_{test} \times p_{result}.$$

The first of these parameters represents the probability that a VL sample will be collected during a clinic visit in which the patient is eligible for VL monitoring according to the current treatment guidelines. The second parameter, p_{test}, represents the probability that the sample is tested and a valid result is obtained. The third parameter, p_{result}, represents the probability that a test result will be returned to the facility and communicated to the patient when they next visit the facility.

[*] Note that this is not necessarily representative of the proportion of ART patients with access to VL monitoring—VL testing was initially implemented at larger facilities, so the proportion of the population with access to VL testing increased much faster than the proportion of facilities offering testing

There is limited data available to estimate appropriate values for these parameters, especially since factors that lead to unsuccessful tests are also likely to be associated with missing data. Various reports on the scale-up of VL monitoring in Malawi and other countries in sub-Saharan Africa indicate that these parameters have changed significantly over time due to the scale-up of VL monitoring and improvements to diagnostic services. For example, [118] reports that the proportion of ART patients in Malawi who had ever received at least one VL test increased from 6% to 51.3% between 2013 and 2018. Annual PEPFAR estimates indicate that in 2018, almost all districts were collecting less than 45% of the targeted number of VL samples [161], while by 2021, all districts were collecting at least 80% of the targeted samples, and many districts collected over 90% of targeted volumes [162].

Based on these data, we modeled each of the three probability parameters as piece-wise linear functions over the years 2013–2028 (see Figure C.4). We considered different rates of increase over the periods 2013–2019 and 2019–2025 to reflect different phases of the VL monitoring scale-up, and assumed that a steady-state was achieved in 2025. We also modeled a significant drop in the probability of VL success during the first 18 months of the Covid-19 pandemic, to account for disruptions in ART clinic services, reagent shortages, and prioritization of Covid-19 samples.

Values for the first two parameters (p_{sample}, p_{test}) were calibrated based on the volume of samples and successful VL tests reported in the national laboratory database between 2013 and 2021 (see Figure C.3), while the probability of result communication (p_{result}) was calibrated to match various statistics reported in the literature (e.g., [29] reported that an average of 27% of centralized VL monitoring tests had missing results in studies across various sub-Saharan African countries in 2016-2019, and [62] found that only 82% of centralized results were delivered in a study in South Africa in 2017).

POC Tests. Empirical data suggests that success rates are significantly higher for POC or near-POC tests than centralized tests. For example, [29] reported success rates of 91% for near-POC testing and 67% for centralized testing, [62] found that 99% of POC results were delivered to patients in South Africa, and [77] reported that

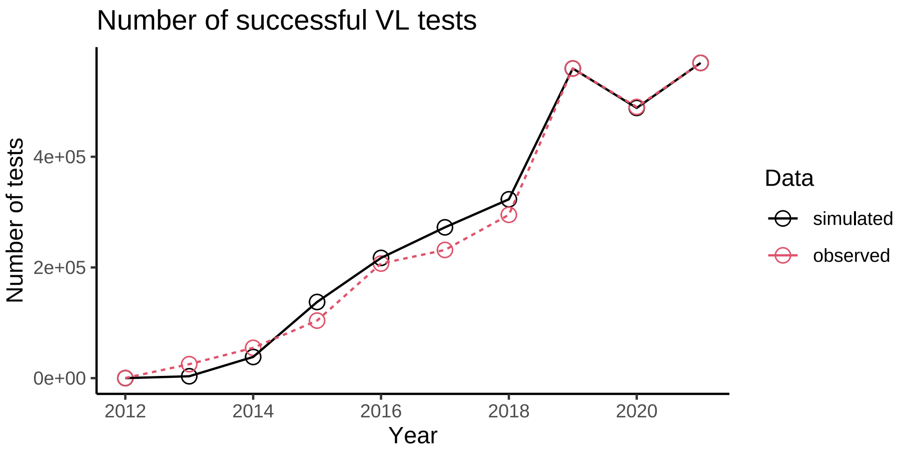


Figure C.3: Simulated vs. observed VL tests.

78% of high VL POC tests had a documented clinical follow-up in a study in Lilongwe, Malawi. However, it is important to consider that many of these studies took place in urban areas and under specialized clinical protocols, which are not necessarily representative of typical conditions in healthcare facilities in Malawi.

For the purposes of our analysis, we assumed that POC testing would initially fail at a similar rate to centralized tests, to account for a period of adjustment to the new technology. We modeled a linear increase in POC success rates during the first two years of the implementation, and assumed that the probability of successful testing and successful result communication would converge to $\bar{p}^P_{test} = \bar{p}^P_{result} = 0.98$ from 2025 onwards. For comparison, the corresponding average success rates for centralized testing were $\bar{p}^C_{test} = \bar{p}^C_{result} = 0.93$. For near-POC results, we assumed that the test success rate would be approximately halfway between that of POC and centralized testing.

C.2.3.3 Turnaround Time for Result Communication and Follow-Up Care

For centralized results, we assumed that result communication and follow-up care would occur at the patient's next appointment, approximately three months after sample collection. Turnaround times for centralized testing have often exceeded this time frame in the past [29], but it is reasonable to assume that efforts to scale-up and improve centralized diagnostic testing should provide results within 2-3 months.

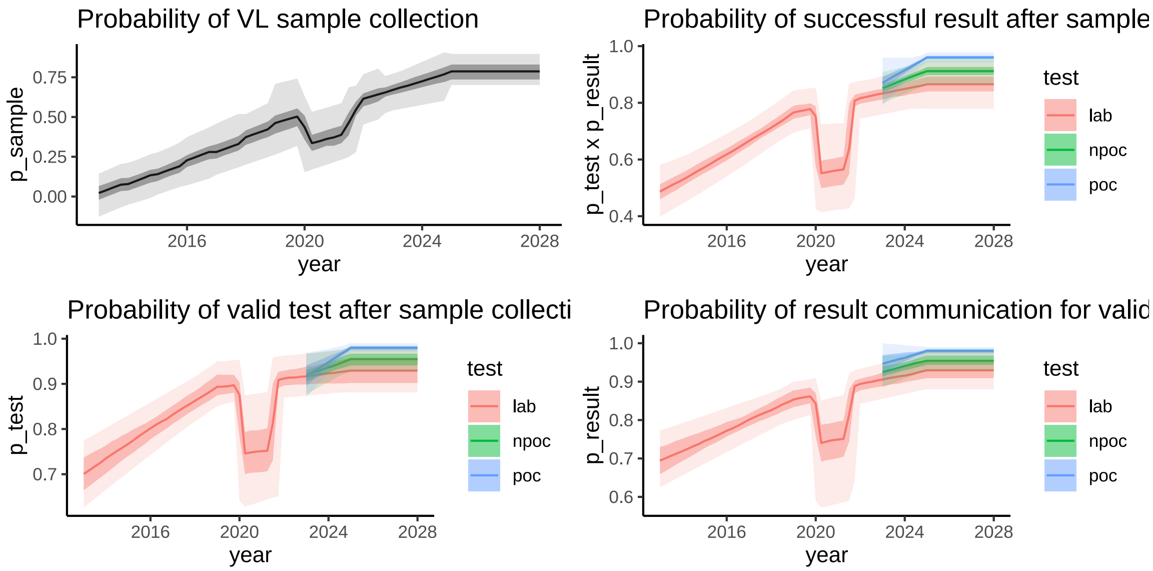


Figure C.4: A summary of the average VL probabilities of sample collection, testing, and result communication over time. The lighter shading represents the full range of values for each parameter, while the darker shading represents the interquartile range.

POC. One of the key benefits of POC testing is shorter turnaround times for test results, which in turn allow faster follow-up care. This assumption is supported by several clinical studies including the STREAM trial, which reported that 99% of POC results were communicated on the same day [62], and a POC study in Malawi which reported 88% of results communicated on the same day [149]. Based on existing literature, we observed that the expected time to result communication varied, with some studies prioritizing same-day results, while other studies required patients to return within a short time frame ($<$ 1 week) to collect results [77]. Given that the simulation model operates in increments of three months, we assumed that small delays of up to a week would have little impact on patient outcomes.

Near-POC Testing. For near-POC testing, same-day results are unlikely due to the need for the sample to be transported to another facility for testing, but it is generally assumed that near-POC results will be communicated to the patient in a relatively short time frame. This often requires patients to return to the facility within a week or two of their original appointment, although it is likely that other communication options (e.g., phone or test notifications) will play a larger role if near-POC testing is widely implemented.

In the absence of rigorous data on follow-up rates for near-POC results, we assumed that approximately 80% of near-POC results would be communicated in a similar time frame to POC tests, while the remaining 20% of results would be communicated in a similar time frame to centralized results (i.e., at the next routine appointment).

C.2.3.4 Response to Adherence Interventions

In the original HIV Synthesis simulation model, the treatment adherence counseling is provided after a patient's first high VL result. The effectiveness of adherence counseling is determined by a global parameter adh_effect_of_meas_alert $= 1 - \rho$, which gives the probability that a patient will change their behavior in response to the intervention. Among patients who respond, 40% of patients permanently increase their treatment adherence to the highest level ($> 90\%$) and 60% temporarily increase their adherence for 6 months. It is assumed that each patient receives only one session of adherence counseling.

In Malawi, treatment adherence counseling is recommended whenever there is a reason to believe that the patient may not be taking their treatment correctly. Our analysis therefore assumes that patients can receive repeated adherence interventions up to once every 12 months, and within 3 months of a high VL result. We maintain the assumption that some patients will permanently increase their treatment adherence after the first intervention (with probability $0.4 \times (1 - \rho)$). For the remaining patients, we assume that a temporary increase in adherence may occur after each adherence intervention with probability $(1 - \rho)$.

While it is difficult to directly measure the effectiveness of adherence counseling, several studies note that many patients with high VL will re-suppress without changing treatment, and that improved adherence is the most likely explanation for this outcome. Empirical studies such as the STREAM trial have reported increased retention of patients receiving POC testing (92%, vs. 82% for centralized testing), which is likely correlated with improved treatment adherence. A recent clinical trial in Haiti found higher average self-reported adherence rates and drug levels among patients who received POC VL monitoring, although the results were not statistically significant (likely due to the relatively small sample size, $n = 150$). Qualitative feedback from both patients and clinicians also indicates that patients value immediate feedback on whether their treatment is working [141], and it is reasonable to assume that discussing adherence issues with patients is more effective if the discussion occurs soon after the issues arise.

We therefore assume that patients receiving POC results will respond to adherence counseling with a higher probability, $(1 - \rho/2)$. For near-POC testing, a patient's response to adherence counseling depends on how quickly they receive their results. In particular, patients who receive their results in a similar time frame to POC testing are assumed to respond at the higher rate, while those who receive their results at the next routing appointment are assumed to respond at the lower rate.

C.2.3.5 Sensitivity Analysis

As illustrated in Figures C.5–C.7, the performance of POC and near-POC testing was sensitive to the assumptions made about failure rates, turnaround times, and adherence impact. Longer delays in returning results had a stronger effect on DALYs, higher rates of failed tests had a significant impact on cost, and more effective adherence counseling improved both costs and DALYs.

C.2.3.6 Priority Criteria

The criteria used to select patients for POC testing in the simulation model are described in Table C.5. At each iteration, POC tests were allocated each priority group in the order that they appear in the table, until no further POC capacity was available. If there was excess capacity after testing all priority patients, the remaining POC tests were randomly allocated to other patients requiring VL monitoring.

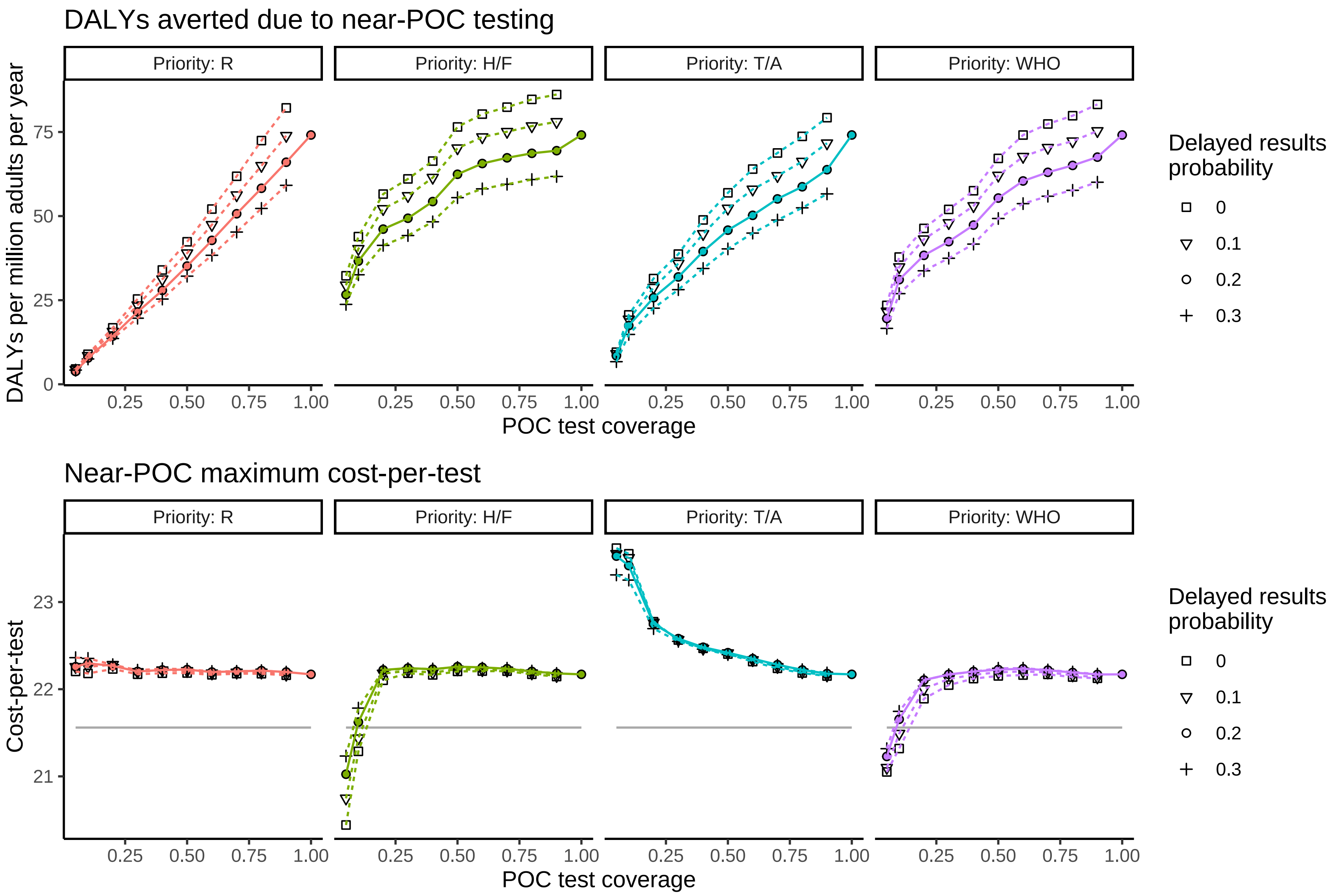


Figure C.5: Sensitivity analysis on the effect of delayed VL results. The delay parameter indicates the proportion of POC results that were returned at the next routine appointment (i.e., 3 months after the test).

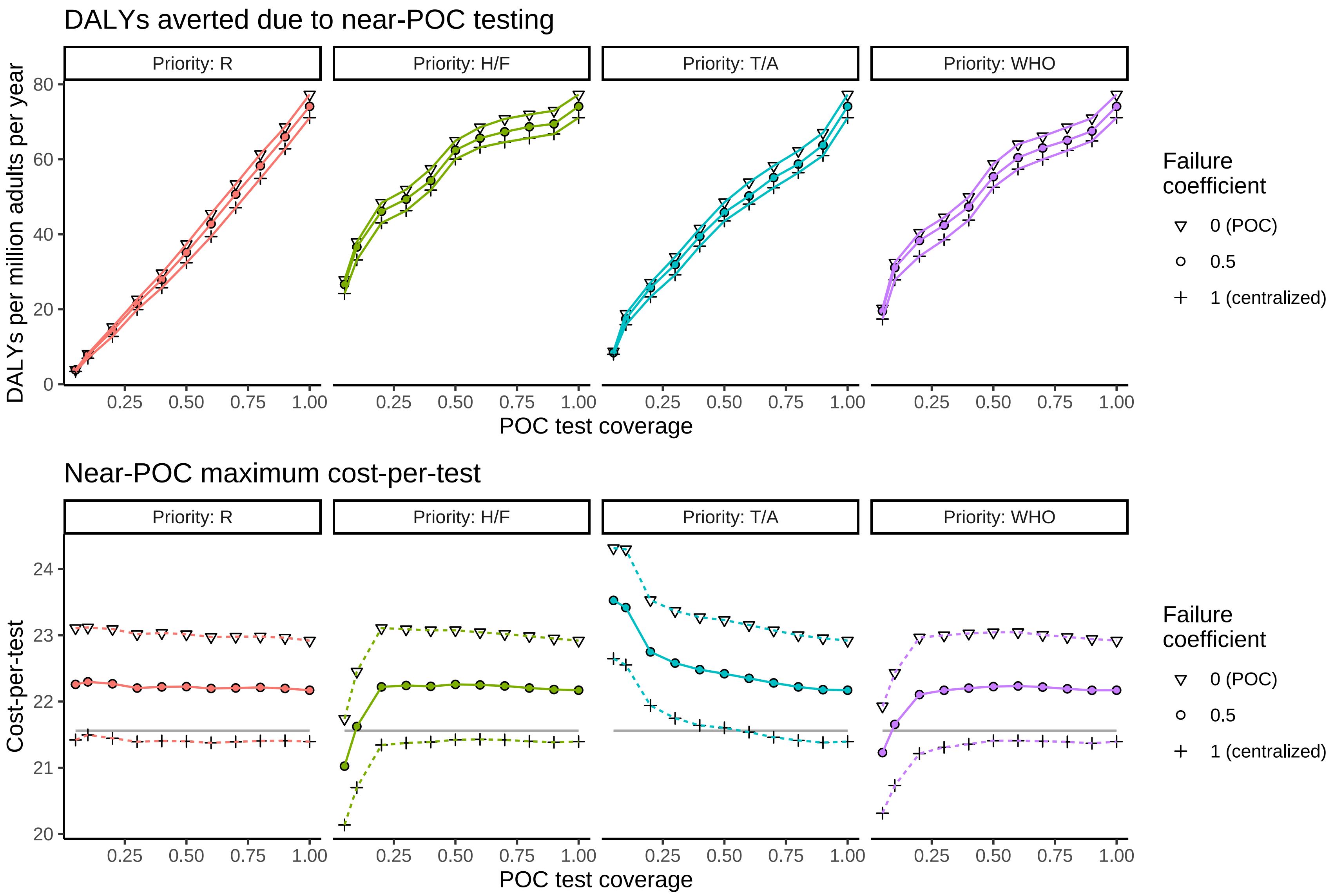


Figure C.6: Sensitivity analysis on the effect of failed VL tests. The failure parameter indicates whether the test failure rate was more similar to the range assumed for POC tests (0), centralized tests (1), or halfway between these rates (0.5).

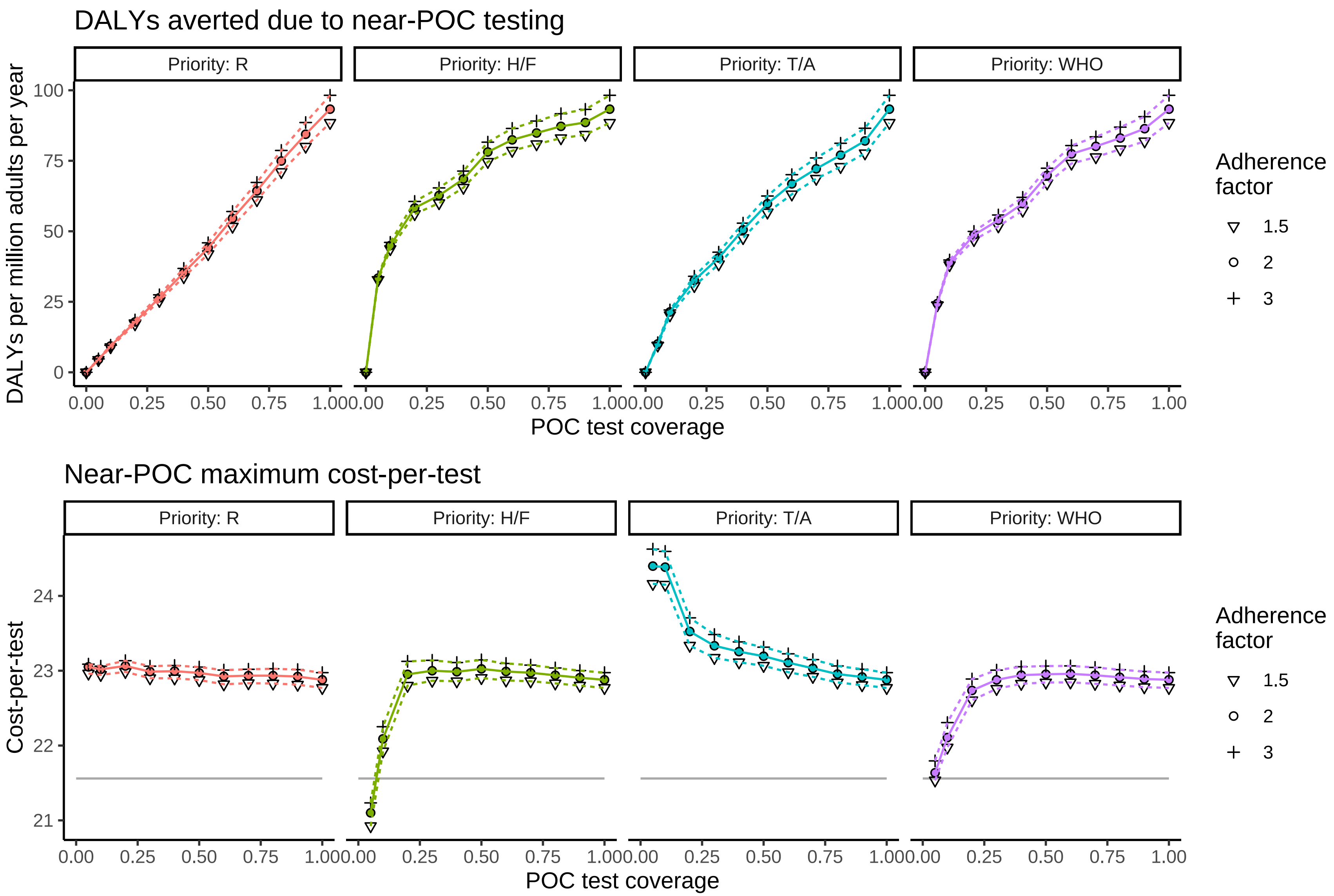


Figure C.7: Sensitivity analysis on the effect of different response rates for treatment adherence counseling. The adherence parameter (α) represents the denominator in the equation adh_effect_of_meas_alert $= 1 - \rho/\alpha$. Higher values of α correspond to more effective adherence counseling after a POC result.

Table C.5: A summary of priority criteria for targeted POC testing of high-risk patients

	WHO	H/F	T/A
Priority 1	pregnancy	confirmatory tests	pregnancy
Priority 2	younger than 20	previous VL result was high	< 1 year on current treatment
Priority 3	confirmatory tests	high VL in the last 12 months	TB or other infections
Priority 4	TB or other infections	TB or other infections	> 1 unprotected sexual partner
Priority 5	first VL test on new treatment		younger than 30

C.3 Optimization

C.3.1 Notation

As in previous sections, we represent healthcare facilities using the subscript i and POC devices using the subscript j. In the model formulation we use the following colors and fonts to distinguish between different types of variables and input data:

- Optimization decision variables
- Operational strategies input data
- Simulation estimates
- POC instruments and facility data
- Model parameters

C.3.2 Decision Variables and Data

We use a set of binary decision variables $\mathrm{d}_{i,j}$ to indicate whether facility i is allocated POC instrument j. For each device and facility, we provide the following input data:

- $\mathrm{Capacity}_j$ — the total number of tests that could be conductted on instrument j in a 5-year period, assuming 7.5 hours of operation 250 days per year.
- $\mathrm{FixedCost}_j$ — the total fixed costs associated with device j over a 5-year period.
- $\mathrm{VarCost}_j$ — the variable costs per test on device j.
- $\mathrm{BASEcoverage}_{i,j}$ — the proportion of VL samples at facility i that could be tested using the capacity of device j, assuming that the facility maintains similar clinic hours to the baseline data from 2021.
- $\mathrm{EXPcoverage}_{i,j}$ — the proportion of VL samples at facility i that could be tested using the capacity of device j, assuming that the facility expands their ART clinic hours to offer VL testing on every weekday.

For convenience, we let instrument $j = 0$ be the baseline scenario (no POC testing, zero capacity, zero fixed and variable costs), while devices $j = 1, \ldots, m$ represent the instruments described in Appendix C.1.1. We also include additional device options $j = m + 1, \ldots, 2m, 2m + 1, \ldots, 3m, 3m + 1, \ldots, 4m$ to represent the same set of instruments with either 25%, 50%, or 75% cost sharing with other disease programs (i.e., devices 1, $m + 1$, $2m + 1$, and $3m + 1$ each represent the same POC instrument, but with reduced fixed costs and testing capacity corresponding to each level of cost-sharing). Finally, we include a near-POC device option, $j = 4m + 1$, which has a fixed cost of \$200 per year (to cover training and administrative overhead associated with implementing near-POC testing).

We use an additional set of binary decision variables $\mathrm{x}_{i,k}$ to represent operational choices for implementing POC testing at each facility. Each option k represents a combination of the following operational decisions:

- $\mathrm{NPOC}_k \in \{0, 1\}$ — whether testing will be performed at POC (0) or near-POC (1)
- $\mathrm{POChub}_k \in \{0, 1\}$ — whether the facility will test near-POC samples referred from other clinics ($\mathrm{POChub}_k = 1$) or not ($\mathrm{POChub}_k = 0$)
- $\mathrm{PropPOC}_k \in \{0.05, 0.1, 0.2, \ldots, 0.9, 1\}$ — the proportion of samples that will be tested at (near-)POC.
- $\mathrm{Priority}_k \in \{1, 2, 3, 4\}$ — which priority strategy will be used to determine which patients have access to limited POC testing capacity.
- $\mathrm{addHours}_k \in \{0, 1\}$ — whether clinic schedules will be expanded to allow testing every day of the week (1) or remain similar to 2021 (0).
- $\mathrm{VarCost}_k$ — the total variable cost for each (near-)POC test, including test cartridges and transportation.
- $\mathrm{LabCost}_{i,k}$ — the total variable cost for each centralized test, including sample transportation (we assume that this value is constant across all operational

strategies k, but varies across facilities i due to different transportation costs).

Note that in the optimization model formulation, $\mathrm{x}_{i,k}$ are decision variables which enumerate all the possible combinations of different operational choices, while NPOC_k, $\mathrm{Priority}_k$, etc. are fixed coefficients that describe the operational strategies associated with the decision variables. We can formulate equivalent models in which these values are treated as decision variables (in which case the the variables $\mathrm{x}_{i,k}$ are unnecessary), but this approach requires non-linear constraints to represent the combined effect of different operational decisions. We avoid this problem by combining all of these decisions into a single set of binary variables.

For each facility i, we use the simulation data to estimate the costs and impact associated with implementing each operational strategy k. The following estimates are required as input to the optimization model:

- $\mathrm{LABtests}_{i,k}$ — the estimated number of centralized tests required over a 5-year period

- $\mathrm{POCtests}_{i,k}$ — the estimated number of (near-)POC tests over a 5-year period.

- $\mathrm{HScosts}_{i,k}$ — the estimated costs of HIV testing and treatment (excluding VL monitoring) over a 5-year period.

- $\mathrm{DALY}_{i,k}$ — the estimated total DALYs averted over a 5-year period.

- $\mathrm{BaseDALY}_i$ — the estimated baseline DALYs averted for 100% centralized testing over a 5-year period.

- $\mathrm{BaseCost}_i$ — the estimated baseline costs for 100% centralized testing over a 5-year period.

In addition to the binary decision variables x and d, we also use the following auxiliary variables in the model formulation:

- daly_i — estimated DALYs averted for facility i

- cost_i — estimated total costs for facility i

- v_i — estimated variable costs for facility i
- f_i — estimated fixed costs for facility i

We denote the districts in Malawi using the index $g = 1, \ldots, 28$, and we let $\mathcal{G}_g$ be the set of all facilities that are in district g. For each district, we define the folowing auxiliary variables:

- $\mathrm{DisCost}_g$ — Total costs in district g
- $\mathrm{DisDALY}_g$ — Total DALYs averted in district g
- $\mathrm{DisPropPOC}_g$ — proportion of samples tested near-POC in district g

C.3.3 Model Constraints

$$\sum_{j=0}^{4m+1} \mathrm{d}_{i,j} = 1 \qquad \forall i \qquad \text{(C.1)}$$

$$\sum_{k=0}^{K} \mathrm{x}_{i,k} = 1 \qquad \forall i \qquad \text{(C.2)}$$

$$\mathrm{daly}_i = \sum_{k=0}^{K} \mathrm{DALY}_{i,k}\mathrm{x}_{i,k} \qquad \forall i \qquad \text{(C.3)}$$

$$\mathrm{v}_i = \sum_{k=0}^{K} \mathrm{x}_{i,k}(\mathrm{VarCost}_k\mathrm{POCtests}_{i,k} + \mathrm{LabCost}_i\mathrm{LABtests}_{i,k} + \mathrm{HScosts}_{i,k}) \quad \forall i \quad \text{(C.4)}$$

$$\mathrm{f}_i = \sum_{j=0}^{4m+1} \mathrm{FixedCost}_j \mathrm{d}_{i,j} \qquad \forall i \tag{C.5}$$

$$\mathrm{cost}_i = \mathrm{f}_i + \mathrm{v}_i \qquad \forall i \tag{C.6}$$

$$\sum_{j=1}^{4m} \mathrm{d}_{i,j} v_j \leq \sum_{k=0}^{K} (1 - \mathrm{NPOC}_k) \mathrm{x}_{i,k} \mathrm{VarCost}_k \qquad \forall i \tag{C.7}$$

$$\sum_{j=1}^{4m} \mathrm{d}_{i,j} \mathrm{BASEcoverage}_{i,j} \geq \sum_{k=0}^{K} \mathrm{x}_{i,k} \mathrm{PropPOC}_k (1 - \mathrm{addHours}_k) \qquad \forall i \tag{C.8}$$

$$\sum_{j=1}^{4m} \mathrm{d}_{i,j} \mathrm{EXPcoverage}_{i,j} \geq \sum_{k=0}^{K} \mathrm{x}_{i,k} \mathrm{PropPOC}_k (\mathrm{addHours}_k) \qquad \forall i \tag{C.9}$$

$$\tag{C.10}$$

Constraints C.1 and C.2 ensure that a single POC device option and a single operational strategy are selected for each healthcare facility.

Constraint C.3 calculates the estimated DALYs averted for each facility based on the amount of POC testing implemented, priority strategy, etc.

Constraint C.4 calculates the variable costs incurred at facility i, including the costs of (near-)POC and centralized testing as well as the health system costs for HIV testing and treatment.

Constraint C.5 calculates the total fixed costs for the POC device or near-POC implementation at facility i. Constraint C.6 calculates the total cost estimate for facility i over a 5-year period.

Constraint C.7 ensures that the variable cost-per-test at facility i is appropriate for the type of device being used (i.e., GeneXpert vs. m-PIMA vs. near-POC).

Constraints C.8 and C.9 require that the proportion of tests conducted at POC are within the capacity of the device allocated to each facility and the decision whether to implement additional clinic days to allow testing on every day of the week.

$$\sum_{j=m+1}^{4m} \mathrm{d}_{i,j} \leq \sum_{k=1}^{K} \mathrm{x}_{i,k}\mathrm{POCHub}_k \qquad \forall i \qquad \text{(C.11)}$$

$$\sum_{k=0}^{K} (\mathrm{NPOC}_k)\mathrm{x}_{i,k} = \mathrm{d}_{i,4m+1} \qquad \forall i \qquad \text{(C.12)}$$

$$\begin{aligned}(\mathrm{cost}_i - \mathrm{BaseCost}_i) \leq 500(\mathrm{daly}_i - \mathrm{BaseDALY}_i) + & \qquad \text{(C.13)}\\ M * \sum_{k=0}^{K} \mathrm{x}_{i,k}(\mathrm{POCHub}_k + \mathrm{NPOC}_k) & \qquad \forall i\end{aligned}$$

$$\mathrm{NPCapacity}_i \leq \sum_{j=1}^{J} \mathrm{d}_{i,j}\mathrm{Capacity}_j - \sum_{k=0}^{K} \mathrm{x}_{i,k}\mathrm{POCtests}_{i,k} \qquad \forall i \qquad \text{(C.14)}$$

$$\mathrm{NPCapacity}_i \leq M \sum_{k=0}^{K} \mathrm{x}_{i,k}\mathrm{POChub}_{i,k} \qquad \forall i \qquad \text{(C.15)}$$

$$\mathrm{NPCapacity}_g = \sum_{i \in \mathcal{G}(g)} \mathrm{NPCapacity}_i \qquad \forall g \qquad \text{(C.16)}$$

$$\mathrm{NPCapacity}_g \geq \sum_{i \in \mathcal{G}_g} \mathrm{x}_{i,k}\mathrm{POCtests}_{i,k} \qquad \forall g \qquad \text{(C.17)}$$

$$\mathrm{DisCost}_g = \sum_{i \in \mathcal{G}_g} \mathrm{cost}_i \qquad \forall g \qquad \text{(C.18)}$$

$$\mathrm{DisDALY}_g = \sum_{i \in \mathcal{G}_g} \mathrm{daly}_i \qquad \forall g \qquad \text{(C.19)}$$

$$(\mathrm{DisCost}_g - \mathrm{DisBaseCost}_g) \leq 500(\mathrm{DisDALY}_g - \mathrm{DisBaseDALY}_g) \qquad \forall g \qquad \text{(C.20)}$$

$$\sum_{k=1}^{K} \mathrm{x}_{i,k}\mathrm{NPOC}_k\mathrm{PropPOC}_k \leq \mathrm{DisPropPOC}_{g(i)} \qquad \forall g \qquad \text{(C.21)}$$

$$\sum_{k=1}^{K} \mathrm{x}_{i,k}(1 + \mathrm{NPOC}_k(\mathrm{PropPOC}_k - 1)) \geq \mathrm{DisPropPOC}_{g(i)} \qquad \forall i \qquad \text{(C.22)}$$

$$\frac{(1 - \mathrm{MaxProp})}{\mathrm{MaxProp}} \sum_{i=1}^{n} \mathrm{x}_{i,k}\mathrm{POCtests}_{i,k} \leq \sum_{i=1}^{n} \mathrm{x}_{i,k}\mathrm{Labtests}_{i,k} \qquad \forall i \qquad \text{(C.23)}$$

$$\sum_{i=1}^{n} f_i \leq \mathrm{FixedBudget} \qquad \forall i \qquad \text{(C.24)}$$

$$\sum_{k=1}^{K} \mathrm{x}_{i,k}\mathrm{Priority}_k = \mathrm{NationalPriority} \qquad \forall i \qquad \text{(C.25)}$$

$$\sum_{k=1}^{K} \mathrm{x}_{i,k}\mathrm{NPOC}_k \leq \mathrm{AllowNPOC} \qquad \forall i \qquad \text{(C.26)}$$

$$\sum_{k=1}^{K} \mathrm{x}_{i,k}?(0 < \mathrm{PropPOC}_k < 1) \leq \mathrm{AllowCombined} \qquad \forall i \qquad \text{(C.27)}$$

$$\sum_{k=1}^{K} \mathrm{x}_{i,k}\mathrm{addHours}_k \leq \mathrm{AllowExpHours} \qquad \forall i \qquad \text{(C.28)}$$

$$\sum_{j=1}^{4m+1} \mathrm{d}_{i,j}\mathrm{DeviceShare}_j \leq \mathrm{MaxShare} \qquad \forall i \qquad \text{(C.29)}$$

Constraint C.11 prohibits cost-sharing for POC devices that will be used for near-POC testing.

Constraint C.12 requires that facilities using near-POC testing are not allocated an instrument, but instead incur the fixed and variable costs associated with near-POC testing.

Constraint C.13 requires that the incremental cost-effectiveness ratio be smaller than 500 at each facility that is allocated a POC device, except for facilities that will also process near-POC tests.

Constraint C.15 and C.14 determine how many near-POC tests can be provided by each POC testing hub, based on the capacity of the device allocated to the facility and the number of POC tests that will be run at the facility. The remaining capacity can be used for near-POC tests, provided that the facility is also selected to be a POC hub.

Constraint C.16 calculates the total near-POC testing capacity in each district, and C.17 requires that the total demand for near-POC testing from facilities in the district should not exceed the available capacity.

Constraint C.18, C.19, and C.20 calculate the total cost and DALY estimates for each district and require that the district ICER does not exceed 500 (note that this constraint is automatically satisfied if all of the individual facilities also meet this threshold, so there is no need to adapt the constraint based on whether or not near-POC testing is used).

Constraint C.21 and C.22 require that if near-POC testing is used, all facilities that do not have their own POC devices must have equal access (proportional to their total sample volumes).

Constraint C.23 restricts the total proportion of VL tests that can be done on (near-)POC instruments.

Constraint C.24 imposes a maximum budget for fixed costs associated with POC instruments and near-POC implementation.

Constraint C.25 requires that all facilities offering POC testing use the same priority strategy to allocate tests to patients (this can be either a fixed strategy, supplied as input data, or a strategy selected during optimization).

Constraints C.26, C.27, C.28, C.29 and the corresponding parameters are used to control which of the operational strategies should be permitted in the optimal solution (e.g., how much cost-sharing is permitted, whether a combination of centralized and POC testing is permitted, etc.).

C.3.4 Solution Approach

The optimization model was coded in Julia [26] using the JuMP [65] package, and solved using Gurobi [89] with default solver parameters. Solving times varied based on the constraints included in the model, but in all cases we were able to find an optimal solution in less than 10 minutes.

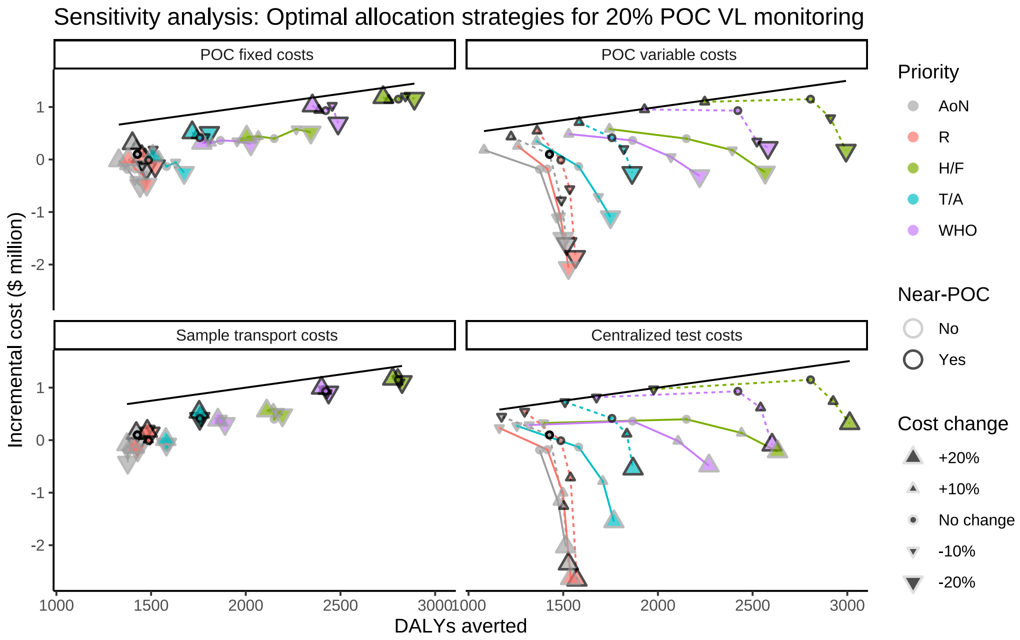


Figure C.8: Sensitivity analysis: a comparison of the optimal POC allocation policies for scenarios with a 10% change in costs associated with POC and/or centralized testing.

C.3.5 Results and Sensitivity Analysis

Figure C.8 shows the results of a sensitivity analysis to investigate the impact of changes in the cost of POC and/or centralized testing on the optimal POC allocation strategies. We considered variations of 10% and 20% in the following costs: POC fixed costs, POC variable costs, centralized test costs, and sample transport costs. In each case we adjusted the input data to the optimization model to reflect the change in costs and then solved the model to find the optimal allocation strategy for that scenario.

The model solutions were most sensitive to changes in POC variable costs and centralized testing costs. This is understandable, given that these costs apply to every single sample tested and account for 70–95% of the total VL monitoring cost. Decreases in POC variable costs and increases in centralized costs resulted in better solutions (i.e., higher DALYs averted and lower costs), as POC testing was more cost-effective in these scenarios. Increases in POC costs or decreases in centralized testing costs resulted in the opposite effect. In scenarios with a 20% increase in POC variable costs or a 20% decrease in centralized costs, no feasible (i.e., cost-effective) solutions were found.

The optimization model was significantly less sensitive to changes in POC fixed costs and sample transport costs, whic contribute a relatively small proportion of the total testing costs

Figure C.9 compares the solutions obtained from the optimization model with different cost-effectiveness constraints. The “strict” constraints required that each POC instrument allocated must have an ICER of at most 500, while the relaxed constraints only required that the total national ICER is below 500. In the later scenario, every optimal solution had a national ICER of exactly $500 per DALY averted, but many of the POC instruments allocated to facilities were significantly less cost-effective. As illustrated in the second row of graphs in Figure C.9, the costs were significantly higher in the relaxed model. These solutions were able to use the DALYs averted due to instruments placed in high-impact locations to offset the cost of allocating instruments to less impactful locations.

The solutions generated in the relaxed model had higher impact (in terms of total DALYs averted) and were also able to cover a wider range of national coverage scenarios. Specifically, the relaxed model was able to construct feasible solutions for $> 60\%$ national coverage without near-POC testing, which was not feasible in the stricter model.

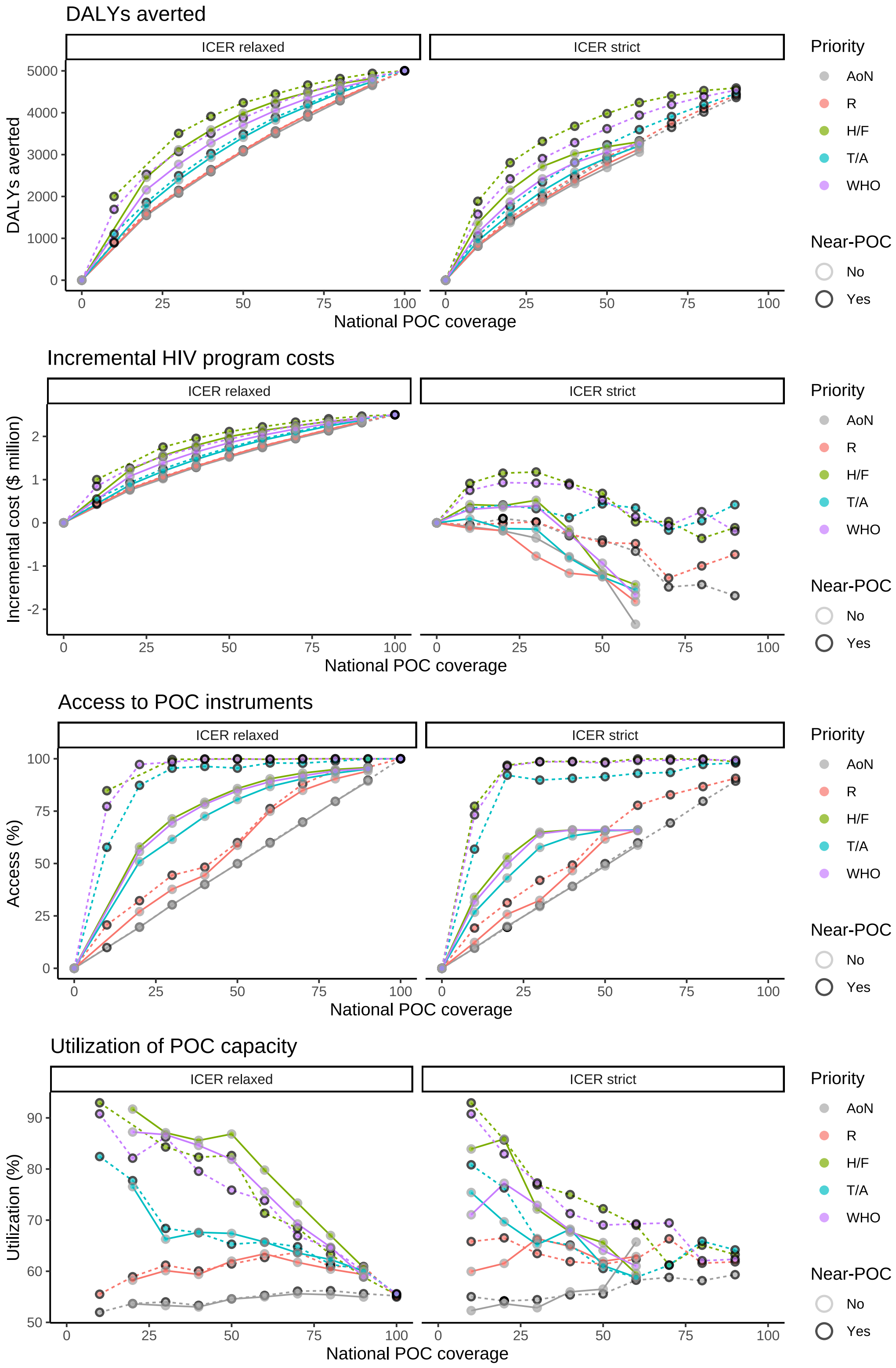


Figure C.9: A comparison of optimal POC allocation strategies for different cost-effectiveness constraints.

Appendix D

Globally Optimized Survival Trees

D.1 Tree Simulations

D.1.1 Tree Generation Algorithm

Algorithm 1 was used to generate ground truth models for simulated datasets.

D.1.2 Survival Distributions

Nodes in simulated trees were randomly assigned one of the following survival distributions:

- Exponential(θ): 0.3, 0.4, 0.6, 0.8, 0.9, 1.15, 1.5, 1.8
- Weibull(k,λ): (0.8,0.4), (0.9,0.5), (0.9,0.7), (0.9,1.1), (0.9,1.5), (1,1.1),(1,1.9), (1.3,0.5)
- Lognormal(μ,σ^2): (0.1,1.0), (0.2,.75), (0.3,0.3), (0.3,0.5), (0.3,0.8), (0.4,0.32), (0.5,0.3), (0.5,0.7)
- Gamma(k,θ): (0.2,.75), (0.3,1.3), (0.3,2), (0.5,1.5), (0.8,1.0), (0.9,1.3), (1.4,0.9), (1.5,0.7)

Algorithm 1 Tree generation algorithm

```
Inputs:
    $X$                                          ▷ $n \times p$ data matrix
    min_bucket                                   ▷ minimum node population
    max_depth                                    ▷ maximum tree depth

procedure INITIALIZE:
    $T \leftarrow \{1\}$                         ▷ list of tree nodes, node 1 is the root node
    status(1) $\leftarrow$ open                  ▷ node status: open nodes may be split, closed nodes are leaf nodes
    population(1) $\leftarrow$ {1,2,...,n}       ▷ observations in each node
    depth(1) $\leftarrow$ 1                      ▷ depth of the node in the tree
procedure GROW TREE:
    while status($k$) = open for any $k \in T$ do
        current_node = select($k \mid k \in T$ and status($k$) = open)   ▷ Select an open node to split
        feature_list = permute(1:$p$+1)          ▷ Randomly order features
        for $j \in$ feature_list do
            if $j = p + 1$ or depth(current_node) = max_depth then
                status(current_node) $\leftarrow$ closed   ▷ Close node without splitting
                break and go to ($A$)
            feature_values = permute(unique($\{X_{ij} \mid i \in$ population(current_node)$\}$))
            for $b \in$ feature_values do        ▷ Attempt to split on feature $j$ with threshold $b$
                $L_1$ = length($\{i \mid i \in$ population(current_node), $X_{ij} \leq b\}$)
                $L_2$ = length($\{i \mid i \in$ population(current_node), $X_{ij} > b\}$)
                if $L_1 \geq$ min_bucket and $L_2 \geq$ min_bucket then   ▷ If split is feasible, create new nodes
                    $T = T \bigcup$ {total_nodes+1,total_nodes+2}
                    status(total_nodes+1) = open
                    depth(total_nodes+1) = depth(current_node)+1
                    population(total_nodes+1) = $\{i \mid i \in$ population(current_node), $X_{ij} \leq b\}$
                    status(total_nodes+2) = open
                    depth(total_nodes+2) = depth(current_node)+1
                    population(total_nodes+2) = $\{i \mid i \in$ population(current_node), $X_{ij} > b\}$
                    status(current_node) $\leftarrow$ closed   ▷ Current node is closed
                    break and go to GROW TREE    ▷ Select another open node to split
    Return $T$
```

D.1.3 Computation Times

Table D.1: Running time summary for the rpart, ctree, and GOST algorithms on synthetically generated data. Across 100 randomized computational experiments, we varied the sample size (n), the number of features (p), and the degree of censoring. The values reflect the average running times of the algorithms in minutes, with the 95% confidence intervals in parenthesis.

n, p	**Censoring**	rpart	ctree	**GOST**
$n = 10$	10%	0.001 (0.001, 0.001)	0.009 (0.008, 0.010)	0.027 (0.024, 0.030)
$p = 1000$	50%	0.001 (0.001, 0.001)	0.010 (0.009, 0.012)	0.033 (0.030, 0.037)
	80%	0.001 (0.001, 0.001)	0.011 (0.010, 0.012)	0.027 (0.024, 0.029)
$n = 50$	10%	0.005 (0.005, 0.005)	0.013 (0.010, 0.015)	0.083 (0.074, 0.091)
$p = 1000$	50%	0.005 (0.005, 0.005)	0.015 (0.012, 0.018)	0.104 (0.093, 0.115)
$n = 100$	80%	0.005 (0.005, 0.006)	0.016 (0.013, 0.019)	0.093 (0.083, 0.104)
$n = 100$	10%	0.008 (0.007, 0.008)	0.022 (0.021, 0.024)	0.149 (0.134, 0.164)
$p = 1000$	50%	0.007 (0.007, 0.008)	0.021 (0.016, 0.027)	0.181 (0.161, 0.202)
	80%	0.007 (0.007, 0.008)	0.023 (0.018, 0.027)	0.178 (0.155, 0.201)
$n = 10$	10%	0.006 (0.005, 0.006)	0.030 (0.027, 0.034)	0.192 (0.169, 0.215)
$p = 5000$	50%	0.006 (0.005, 0.006)	0.034 (0.030, 0.039)	0.188 (0.165, 0.211)
	80%	0.006 (0.005, 0.006)	0.037 (0.035, 0.039)	0.205 (0.187, 0.224)
$n = 50$	10%	0.017 (0.016, 0.018)	0.043 (0.029, 0.057)	0.511 (0.458, 0.563)
$p = 5000$	50%	0.017 (0.016, 0.018)	0.049 (0.031, 0.067)	0.506 (0.443, 0.570)
	80%	0.017 (0.016, 0.019)	0.054 (0.036, 0.072)	0.639 (0.575, 0.704)
$n = 100$	10%	0.033 (0.030, 0.036)	0.077 (0.047, 0.107)	0.895 (0.795, 0.996)
$p = 5000$	50%	0.033 (0.031, 0.035)	0.089 (0.060, 0.119)	0.897 (0.769, 1.025)
	80%	0.033 (0.031, 0.035)	0.110 (0.081, 0.140)	1.172 (1.034, 1.311)
$n = 10$	10%	0.010 (0.009, 0.011)	0.085 (0.076, 0.094)	0.548 (0.487, 0.609)
$p = 10000$	50%	0.010 (0.009, 0.011)	0.089 (0.079, 0.099)	0.478 (0.423, 0.533)
	80%	0.011 (0.010, 0.011)	0.098 (0.087, 0.108)	0.561 (0.513, 0.608)
$n = 50$	10%	0.035 (0.033, 0.038)	0.124 (0.085, 0.163)	1.268 (1.117, 1.419)
$p = 10000$	50%	0.039 (0.037, 0.041)	0.149 (0.114, 0.183)	1.363 (1.221, 1.505)
	80%	0.036 (0.033, 0.038)	0.144 (0.121, 0.167)	1.713 (1.529, 1.896)
$n = 100$	10%	0.061 (0.055, 0.068)	0.121 (0.069, 0.174)	2.275 (1.994, 2.557)
$p = 10000$	50%	0.065 (0.061, 0.068)	0.197 (0.134, 0.260)	2.307 (2.022, 2.592)
	80%	0.062 (0.057, 0.067)	0.227 (0.175, 0.279)	3.110 (2.785, 3.435)
$n = 10$	10%	0.022 (0.020, 0.023)	0.462 (0.414, 0.509)	2.295 (2.026, 2.564)
$p = 25000$	50%	0.021 (0.019, 0.024)	0.486 (0.427, 0.545)	2.587 (2.299, 2.875)
	80%	0.026 (0.025, 0.027)	0.535 (0.460, 0.610)	2.277 (2.082, 2.471)
$n = 50$	10%	0.086 (0.080, 0.092)	0.458 (0.357, 0.559)	4.726 (4.210, 5.242)
$p = 25000$	50%	0.088 (0.080, 0.095)	0.494 (0.365, 0.624)	4.443 (3.845, 5.042)
	80%	0.085 (0.076, 0.095)	0.621 (0.470, 0.771)	5.603 (5.034, 6.171)
$n = 100$	10%	0.171 (0.158, 0.185)	0.826 (0.563, 1.090)	6.575 (5.786, 7.363)
$p = 25000$	50%	0.155 (0.129, 0.182)	0.718 (0.103, 1.539)	8.404 (7.468, 9.340)
	80%	0.140 (0.042, 0.322)	0.710 (2.398, 3.819)	10.28 (9.170, 11.38)
$n = 10$	10%	0.045 (0.041, 0.050)	1.766 (1.616, 1.916)	7.078 (6.274, 7.881)
$p = 50000$	50%	0.049 (0.045, 0.053)	1.827 (1.684, 1.970)	9.112 (8.052, 10.17)
	80%	0.050 (0.045, 0.056)	1.856 (1.696, 2.017)	8.004 (7.092, 8.915)
$n = 50$	10%	0.168 (0.156, 0.180)	1.742 (1.450, 2.035)	13.42 (11.97, 14.87)

Continued on next page

Table D.1 – continued from previous page

n, p	**Censoring**	rpart	ctree	**GOST**
$p = 50000$	50%	0.182 (0.169, 0.194)	1.886 (1.656, 2.116)	15.02 (13.20, 16.83)
	80%	0.186 (0.174, 0.197)	1.993 (1.825, 2.161)	16.25 (14.59, 17.90)
$n = 100$	10%	0.343 (0.308, 0.378)	2.217 (1.620, 2.814)	20.72 (18.46, 22.97)
$p = 50000$	50%	0.327 (0.302, 0.352)	2.223 (1.760, 2.686)	21.79 (18.99, 24.59)
	80%	0.333 (0.291, 0.376)	2.498 (2.079, 2.916)	26.66 (23.87, 29.45)
$n = 10$	10%	0.087 (0.077, 0.098)	6.018 (5.608, 6.429)	23.69 (23.03, 24.34)
$p = 100000$	50%	0.095 (0.085, 0.105)	6.271 (5.799, 6.742)	24.46 (23.90, 25.01)
	80%	0.100 (0.089, 0.111)	6.329 (5.949, 6.709)	21.33 (20.81, 21.89)
$n = 50$	10%	0.377 (0.340, 0.415)	6.491 (5.532, 7.450)	34.15 (33.35, 34.95)
$p = 100000$	50%	0.361 (0.319, 0.402)	6.476 (5.843, 7.108)	37.23 (36.34, 38.12)
	80%	0.386 (0.343, 0.428)	6.769 (6.252, 7.286)	35.14 (34.29, 35.98)
$n = 100$	10%	0.732 (0.640, 0.823)	6.773 (5.529, 8.017)	47.60 (46.22, 48.98)
$p = 100000$	50%	0.760 (0.659, 0.860)	7.670 (6.613, 8.728)	53.03 (51.43, 54.63)
	80%	0.744 (0.659, 0.829)	7.646 (6.602, 8.691)	52.44 (50.88, 53.99)

D.2 Real-World Survival Datasets

D.2.1 Wisconsin Longitudinal Study

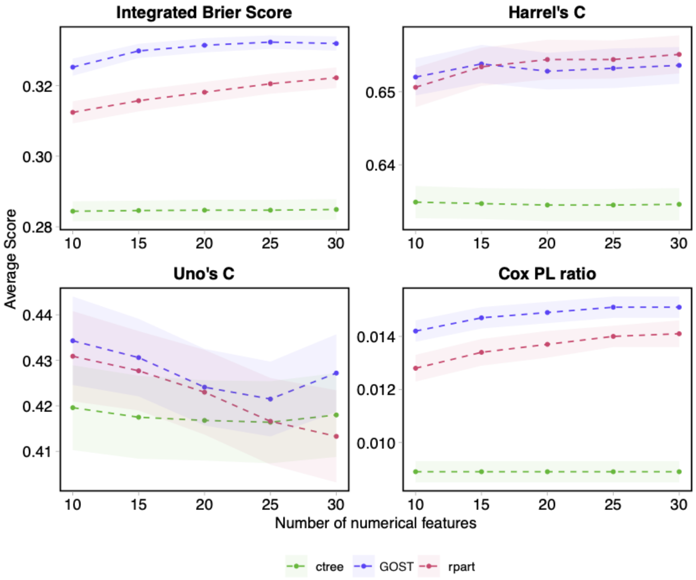


Figure D.1: Average performance of survival tree models on the WLS dataset with different number of numerical variables. The shaded regions represent the 95% confidence intervals across 100 randomized experiments.

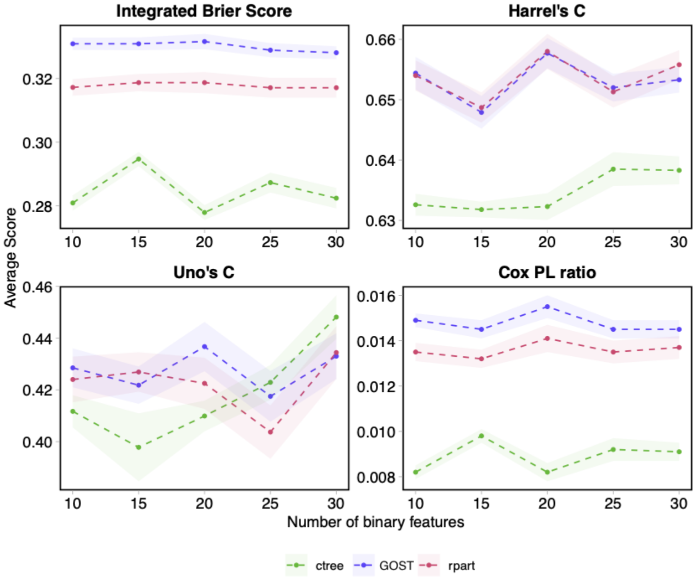


Figure D.2: Average performance of survival tree models on the WLS dataset with different number of binary variables. The shaded regions represent the 95% confidence intervals across 100 randomized experiments.

D.2.2 The Framingham Heart Study

D.2.2.1 FHS Dataset Inclusion Criteria

- Participation in the Original and Offspring cohort of the FHS.
- Formal diagnosis with chd (as indicated by the records of FHS).
- Participants outcomes were followed for 10 consecutive years after diagnosis.

D.2.2.2 Saturated Trees

Using the Framingham Heart Study data, we conducted additional experiments to explore the relationship between the size of the tree and the model's quantative performance. We trained a sequence of fully saturated trees by setting the complexity parameter to zero and varying the maximum depth parameter in each algorithm. Thus, each algorithm returned the best performing tree given two, four, eight, and 16 leaf nodes (see Figures D.3–D.5 for examples). Table D.2 presents the results of our analysis for the three algorithms considered with respect to the IBR, Cox PL, Harrell's C, and Uno's C metrics. We do not report confidence intervals as the same tree was recovered for each set of parameters across all algorithms.

Table D.2: Average scores in 100 randomized experiments for GOST, rpart, ctree models on the FHS dataset for different values of the maximum depth parameter.

Leaf Nodes	**Algorithm**	**IBR**	**Cox PL**	**Harrell's C**	**Uno's C**
2	**ctree**	0.393	0.0016	0.5368	0.5936
	GOST	0.3932	0.0021	0.5429	0.5894
	rpart	0.3914	0.0021	0.5429	0.5894
4	**ctree**	0.3996	0.0029	0.5588	0.585
	GOST	0.4028	0.0031	0.5615	0.5861
	rpart	0.3881	0.0027	0.546	0.5714
8	**ctree**	0.4006	0.0064	0.5645	0.5897
	GOST	0.4041	0.0066	0.567	0.5897
	rpart	0.4031	0.0072	0.5686	0.5871
16	**ctree**	0.3851	0.0069	0.5608	0.5738
	GOST	0.3878	0.0098	0.5713	0.5858
	rpart	0.3566	0.0052	0.5488	0.5577

Given that each algorithm is constrained to produce a tree of the same size and shape, there is less variation in these results than in models that are pruned by a complexity parameter. We note that for the smallest trees (2 nodes), the rpart and GOST results are virtually identical. This outcome is expected since these trees contain a single split, and both algorithms use similar criteria to identify the best split. In larger trees, however, the GOST models performed slightly better than rpart.

In general, GOST had the best average performance across all tree sizes for the IBR and Uno's C metrics and in the majority of cases for the Cox PL and Harrell's C metrics. While these results are not representative of the general performance of the algorithms with cross-validation, they suggest that GOST is able to recover more accurate data partitions when the size of the model is constrained.

These results provide some insight into the tradeoff between complexity and accuracy. A smaller number of leaf nodes is arguably associated with higher model interpretability, as patient profiles can be characterized with fewer features. In this case, models with 8 leaf nodes offer the best average performance by a relatively small margin, and in certain contexts it may be appropriate to select a smaller tree size even if this lowers the model's predictive power.

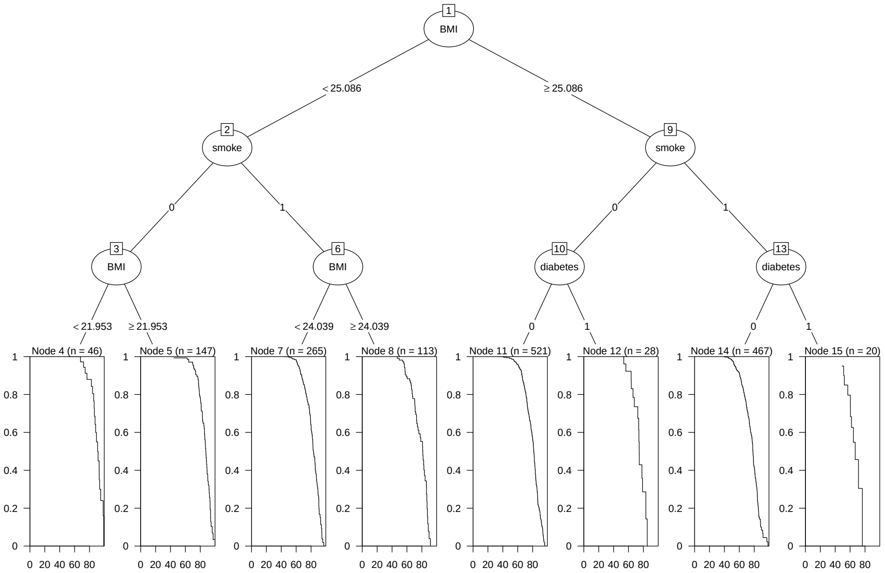


Figure D.3: Saturated tree partition of the FHS dataset using the rpart algorithm when the feature space is restricted to diabetes, smoking, and BMI.

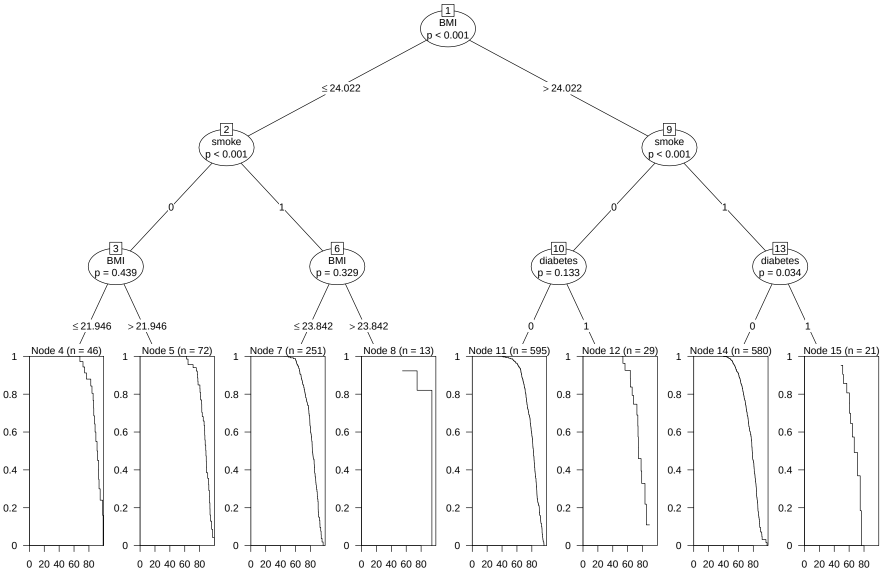


Figure D.4: Saturated tree partition of the FHS dataset using the ctree algorithm when the feature space is restricted to diabetes, smoking, and BMI.

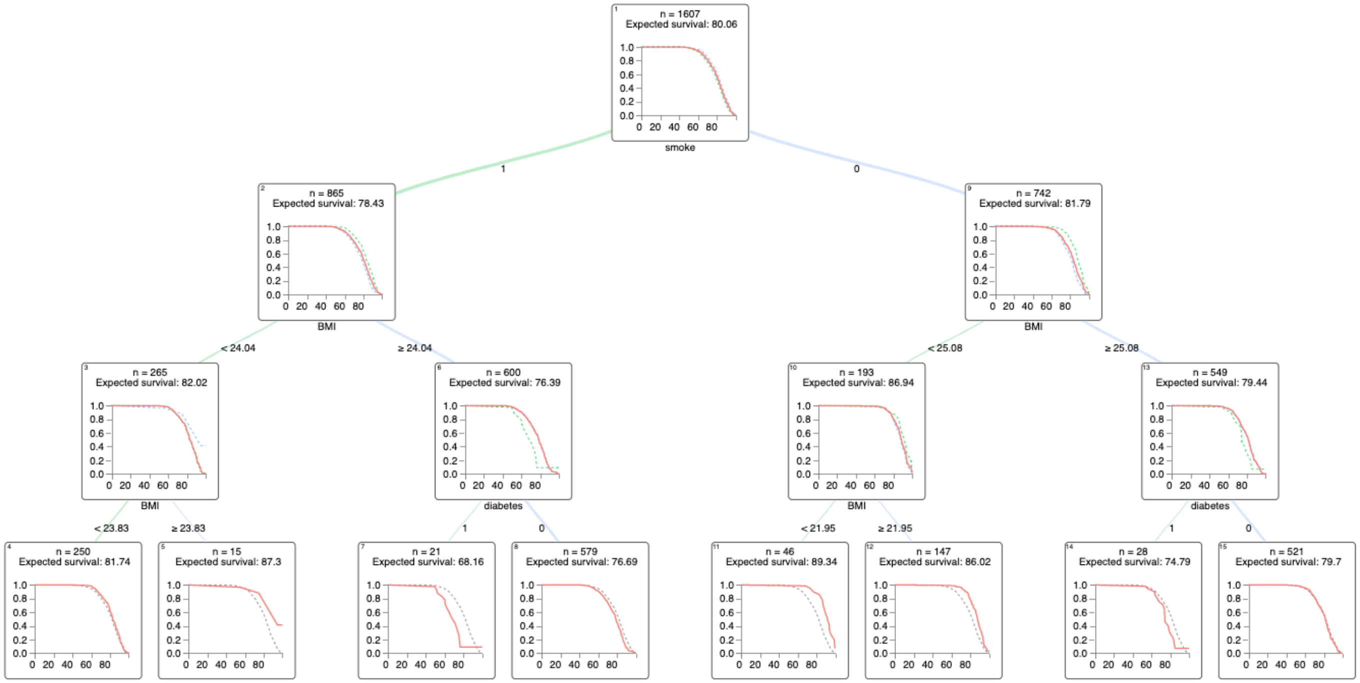


Figure D.5: Saturated tree partition of the FHS dataset using the GOST algorithm when the feature space is restricted to diabetes, smoking, and BMI.

]

D.3 Supplementary Experiments for Real Datasets with Simulated Censoring

This appendix describes supplementary experiments which compare the performance of the GOST, rpart and ctree algorithms on 44 real-world datasets. The datasets used for this analysis were sourced from the UCI repository [63], a well-established resource for the machine learning community. The datasets contained uncensored continuous outcome measures, and we simulated artificial censoring to test the performance of the survival tree algorithms.

The selected datasets* had sample sizes ranging from 63 observations to 100000, and the maximum number of features considered was 383. We used the censoring procedure described in Section 5.6.1 to generate 9 versions of each dataset with different levels of censoring (0%,10%,. . . ,80%), We then split each dataset into training and testing sets (50%) and compared the performance of the three tree algorithms on each dataset.

We applied the 5-fold cross-validation procedure described in Section 5.6.1 to select the depth and complexity of each tree, allowing tree depths of up to 7 (128 leaf nodes). Both the GOST and ctree algorithms produced trees with over 100 leaf nodes in some of the largest datasets, while the largest rpart trees had only 77 nodes. The smaller size of the rpart trees indicates that larger models performed poorly in the cross-validation step.

D.3.1 Summary of Results for UCI Experiments with Simulated Censoring

On average, the GOST models outperformed the other two algorithms across all 5 accuracy metrics. A summary of each algorithm's performance is given in Tables D.3–

*We excluded the following types of datasets from our analysis: (1) datasets used for time series predictions (multiple observations of each individual); (2) datasets with unclear variable definitions; (3) datasets which required significant cleaning, pre-processing, or recoding; (4) datasets with too many variables (p) to cross-validate all three algorithms in reasonable times. Dataset selection was independent of the analysis of model accuracy.

D.4 and Figure D.6, and aggregated results for each dataset are displayed in Table D.6. The difference in performance was not statistically significant for the Cox ratios and Harrell's C scores, where all three algorithms had very similar average outcomes, but GOST models did score significantly better than the other algorithms on the remaining metrics. GOST models achieved the best score in 48-60% of the datasets tested, while the other algorithms each had undominated scores in 27-39% of datasets.

Table D.3: Average scores for GOST, rpart and ctree models on real-world datasets. The final columns show the one-sided p-values for paired t-tests comparing the outcome metrics on each dataset.

	Mean score			Paired T-Test H_1:	
	GOST	rpart	ctree	$S_{GOST} > S_{rpart}$	$S_{GOST} > S_{ctree}$
Cox Ratio	**0.1118**	0.1091	0.1090	p=0.2288	p=0.2222
Harrell's C	**0.7873**	0.7866	0.7818	p=0.4355	p=0.1045
Uno's C	**0.6650**	0.6523	0.6441	**p=0.0288**	**p=0.0013**
Brier Point Ratio	**0.3841**	0.3627	0.3516	**p=0.0001**	**p$< 10^{-5}$**
Intg. Brier Ratio	**0.4451**	0.4262	0.4231	**p=0.0135**	**p=0.0055**

Table D.4: The percentage of datasets for which each algorithm was undominated by the other algorithms. Note that rows do not sum to 100, as several datasets were tied.

	GOST	rpart	ctree
Cox Ratio	48.7	32.8	36.4
Harrell's C	57.3	30.8	33.6
Uno's C	59.3	27.3	34.1
Brier Point Ratio	56.6	33.3	38.4
Intg. Brier Ratio	57.6	30.6	33.6

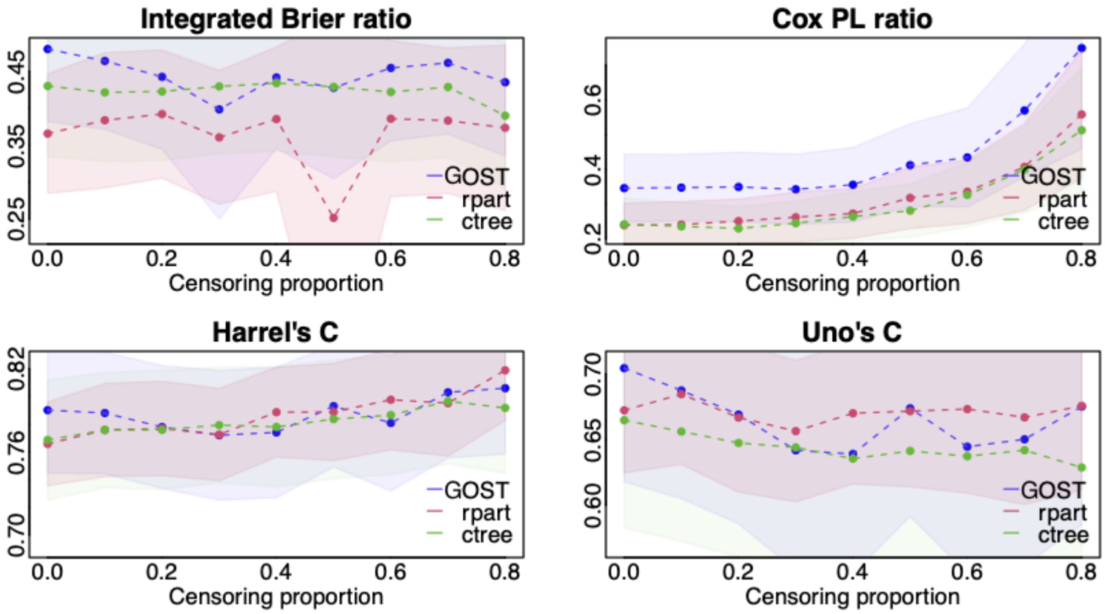


Figure D.6: Average performance of survival tree models on real datasets with different levels of censoring. Confidence intervals are large due to the significant variability between datasets. However matched pairs analysis yields statistically significant results.

D.3.2 Nemenyi Critical Diagrams for UCI Experiments with Simulated Censoring

In this section, we show Nemenyi Critical Diagrams for the results from the UCI datasets for different levels of censoring [70]. These graphs highlight statistically significant differences in the overall rankings of GOST, rpart, and ctree. To generate the diagrams, first the Friedman Rank Test was performed to compare the relative performance of the algorithms for a given level of censoring across all datasets. Second, the Wilcoxon-Holm method was performed to detect pairwise significance. In each diagram, every survival analysis method is plotted according to its average relative rank on a number line from one to three. If the methods are not associated with statistically significant differences in their overall rankings, they are joined by a horizontal line. Figures D.7-D.11 show the Nemenyi Critical Diagrams comparing the performance of the three algorithms for all evaluation metrics considered. This analysis is in agreement with the results presented in Sections 5.6–5.7. We demonstrate that GOST provide a consistent improvement of accuracy across all metrics compared to the other two algorithms. The relative performance of the method improves in instances with a higher sample size.

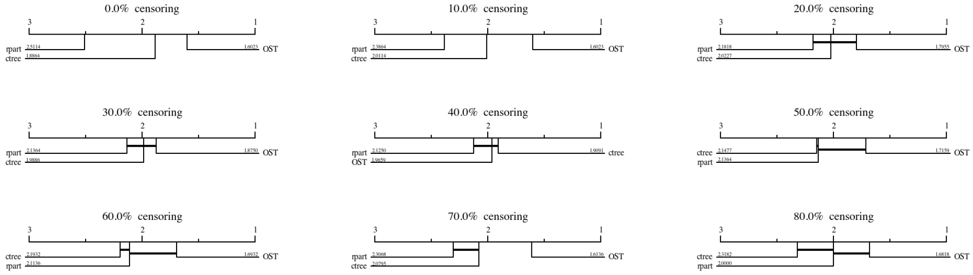


Figure D.7: Nemenyi critical diagrams comparing the relative ranking of all methods with respect to the Integrated Brier Score metric.

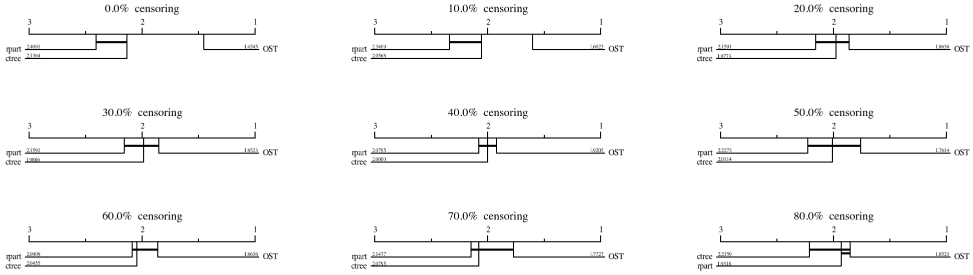


Figure D.8: Nemenyi critical diagrams comparing the relative ranking of all methods with respect to the Harrell's C Score metric.

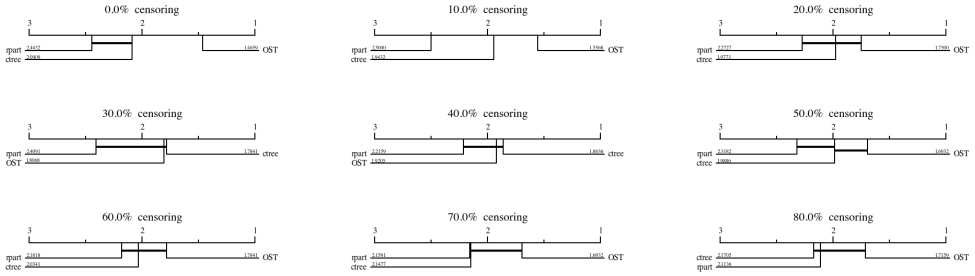


Figure D.9: Nemenyi critical diagrams comparing the relative ranking of all methods with respect to the Uno's C Score metric.

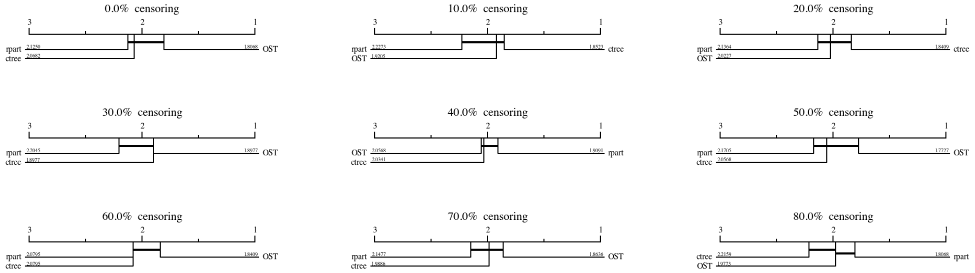


Figure D.10: Nemenyi critical diagrams comparing the relative ranking of all methods with respect to the Cox Partial Likelihood metric.

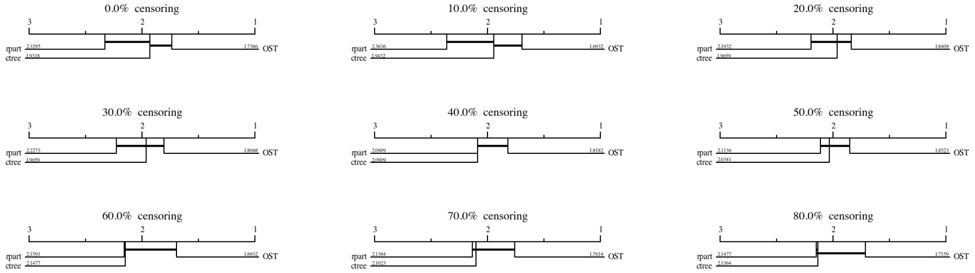


Figure D.11: Nemenyi critical diagrams comparing the relative ranking of all methods with respect to the Brier Point Ratio metric.

D.3.3 Aggregate Results for UCI Experiments with Simulated Censoring

Table D.5: Average scores for GOST, rpart, ctree models for real world datasets for each level of censoring.

Censoring & Method		**Integrated Brier**		**Harrel's C Score**		**Uno's C Score**		**Cox Partial Likelihood**		**Brier Point Ratio**	
0	ctree	0.43	(0.047)	0.769	(0.022)	0.665	(0.041)	0.097	(0.016)	0.325	(0.068)
	GOST	**0.48**	(0.049)	**0.79**	(0.023)	**0.704**	(0.043)	**0.1**	(0.018)	**0.366**	(0.07)
	rpart	0.404	(0.042)	0.767	(0.018)	0.654	(0.035)	0.084	(0.011)	0.332	(0.055)
0.1	ctree	0.422	(0.046)	0.776	(0.021)	0.656	(0.042)	**0.096**	(0.015)	0.323	(0.067)
	GOST	**0.464**	(0.046)	**0.788**	(0.022)	**0.687**	(0.041)	0.087	(0.012)	**0.352**	(0.067)
	rpart	0.403	(0.043)	0.767	(0.02)	0.636	(0.04)	0.075	(0.01)	0.323	(0.053)
0.2	ctree	0.423	(0.046)	0.776	(0.021)	0.647	(0.043)	0.091	(0.014)	0.313	(0.067)
	GOST	**0.443**	(0.049)	0.778	(0.022)	**0.669**	(0.041)	**0.097**	(0.014)	**0.349**	(0.068)
	rpart	0.421	(0.045)	**0.781**	(0.019)	0.66	(0.038)	0.093	(0.013)	0.339	(0.066)
0.3	ctree	0.43	(0.045)	0.779	(0.021)	0.644	(0.042)	0.097	(0.013)	0.316	(0.064)
	GOST	0.399	(0.074)	0.772	(0.023)	0.642	(0.045)	**0.1**	(0.015)	**0.331**	(0.065)
	rpart	**0.434**	(0.045)	**0.784**	(0.02)	**0.659**	(0.038)	0.094	(0.014)	0.315	(0.063)
0.4	ctree	0.434	(0.046)	0.778	(0.022)	0.635	(0.044)	0.097	(0.013)	0.307	(0.065)
	GOST	**0.442**	(0.048)	0.774	(0.023)	0.639	(0.045)	**0.105**	(0.015)	**0.346**	(0.066)
	rpart	0.429	(0.049)	**0.784**	(0.021)	**0.646**	(0.042)	0.104	(0.015)	0.332	(0.066)
0.5	ctree	0.429	(0.047)	0.784	(0.021)	0.641	(0.042)	0.109	(0.014)	0.387	(0.055)
	GOST	0.428	(0.061)	**0.793**	(0.022)	**0.674**	(0.041)	**0.12**	(0.017)	**0.418**	(0.056)
	rpart	**0.443**	(0.046)	0.782	(0.022)	0.635	(0.043)	0.108	(0.015)	0.384	(0.055)
0.6	ctree	0.422	(0.047)	0.787	(0.021)	0.637	(0.043)	0.114	(0.014)	0.434	(0.057)
	GOST	**0.455**	(0.049)	0.781	(0.024)	0.645	(0.048)	**0.118**	(0.016)	**0.456**	(0.058)
	rpart	0.439	(0.047)	**0.791**	(0.022)	**0.648**	(0.043)	**0.118**	(0.015)	0.436	(0.057)
0.7	ctree	0.429	(0.047)	0.797	(0.022)	0.642	(0.044)	0.136	(0.02)	0.404	(0.091)
	GOST	**0.461**	(0.048)	0.803	(0.024)	0.65	(0.048)	0.132	(0.02)	**0.438**	(0.092)
	rpart	0.439	(0.046)	**0.807**	(0.021)	**0.668**	(0.04)	**0.138**	(0.018)	0.425	(0.09)
0.8	ctree	0.39	(0.049)	0.792	(0.023)	0.629	(0.046)	0.145	(0.023)	0.357	(0.106)
	GOST	**0.435**	(0.05)	0.806	(0.024)	**0.675**	(0.045)	0.146	(0.027)	**0.402**	(0.108)
	rpart	0.424	(0.047)	**0.816**	(0.021)	0.667	(0.043)	**0.168**	(0.023)	0.378	(0.105)

Table D.6: Average scores for GOST, rpart, ctree for each dataset across all levels of censoring.

Dataset	**Method**	Integrated Brier	Harrell's C	Uno's C	Cox PL	Brier point
3D Spatial Network [103]	GOST	**0.44**	**0.82**	**0.79**	**0.05**	**0.48**
$n = 100000$	rpart	0.33	0.77	0.73	**0.05**	0.35
$p = 1$	ctree	0.39	0.79	0.76	**0.05**	0.41
Airfoil Self Noise [63]	GOST	**0.39**	**0.83**	**0.77**	**0.09**	**0.53**
$n = 1503$	rpart	0.33	0.78	0.7	**0.09**	0.42
$p = 4$	ctree	0.35	0.78	0.71	0.08	0.46
Appliances Energy Prediction [40]	GOST	**0.19**	**0.74**	**0.7**	**0.03**	**0.14**
$n = 19735$	rpart	0.18	0.73	0.69	**0.03**	0.13
$p = 25$	ctree	0.18	**0.74**	**0.7**	**0.03**	0.12
Automobile [40]	GOST	0.03	0.53	0.08	0.01	0
$n = 164$	rpart	**0.07**	**0.65**	**0.41**	**0.05**	**0.11**
$p = 23$	ctree	0.06	0.61	0.27	0.03	**0.11**
Auto MPG [63]	GOST	0.55	0.85	**0.79**	0.19	0.58
$n = 398$	rpart	**0.56**	**0.87**	0.78	0.2	**0.6**
$p = 7$	ctree	0.55	**0.87**	0.77	**0.21**	0.58
Behavior Urban Traffic	GOST	0.18	0.66	0.37	0.08	0.13
$n = 135$	rpart	**0.2**	**0.67**	**0.41**	**0.09**	**0.16**
$p = 16$	ctree	0.18	0.64	0.33	0.08	0.14
Bike Sharing	GOST	0.92	**0.98**	**0.96**	0.15	0.94
$n = 17379$	rpart	0.88	0.96	0.93	0.09	0.91
$p = 13$	ctree	**0.93**	**0.98**	**0.96**	**0.2**	**0.95**
Blog Feedback [68]	GOST	**0.39**	0.84	0.79	**0.03**	0.17
$n = 52397$	rpart	**0.39**	**0.85**	0.8	**0.03**	**0.18**
$p = 279$	ctree	0.38	**0.85**	**0.82**	**0.03**	0.17
Buzz in Social Media [104]	GOST	**0.77**	**0.92**	**0.91**	**0.13**	**0.76**
$n = 100000$	rpart	0.75	0.91	0.88	0.12	0.74
$p = 76$	ctree	**0.77**	**0.92**	0.9	0.12	0.75
Cargo2000 [133]	GOST	**1**	**1**	**1**	**0.21**	0.22
$n = 3943$	rpart	**1**	**1**	**1**	**0.21**	**0.23**
$p = 95$	ctree	0.84	0.95	0.9	0.16	0.17
Communities Crime [185]	GOST	0.64	0.89	0.81	0.17	0.68
$n = 2215$	rpart	0.65	0.89	0.83	0.17	0.7
$p = 145$	ctree	**0.69**	**0.91**	**0.85**	**0.19**	**0.75**
Computer Hardware [63]	GOST	**0.69**	**0.86**	**0.74**	0.24	**0.73**
$n = 209$	rpart	0.61	0.83	0.67	0.27	0.68
$p = 8$	ctree	0.65	0.85	0.7	**0.29**	0.62
Concrete Slump [236]	GOST	0.07	0.62	0.27	0.04	0.11
$n = 103$	rpart	**0.14**	**0.66**	**0.35**	**0.07**	**0.13**
$p = 6$	ctree	0.03	0.56	0.14	0.02	0.05
Concrete Strength [235]	GOST	**0.42**	**0.84**	0.74	0.11	0.5
$n = 1030$	rpart	0.41	0.83	0.74	**0.13**	0.47

Continued on next page

Table D.6 – continued from previous page

Dataset	Method	Integrated Brier	Harrell's C	Uno's C	Cox PL	Brier point
$p = 7$	ctree	0.4	0.82	**0.75**	0.12	**0.51**
CSM [5]	GOST	0.25	0.71	0.48	0.08	0.34
$n = 232$	rpart	**0.32**	**0.76**	0.56	**0.11**	**0.42**
$p = 11$	ctree	0.25	0.73	**0.57**	0.09	0.26
Cycle Power	GOST	**0.73**	**0.92**	**0.89**	0.16	**0.75**
$n = 9568$	rpart	0.71	0.91	0.86	0.17	0.72
$p = 3$	ctree	**0.73**	**0.92**	**0.89**	**0.18**	**0.75**
Electrical Stability [63]	GOST	**0.4**	**0.82**	**0.79**	**0.08**	**0.44**
$n = 10000$	rpart	0.34	0.79	0.75	0.06	0.37
$p = 11$	ctree	0.39	**0.82**	**0.79**	**0.08**	**0.44**
Energy efficiency 1 [210]	GOST	**0.95**	**0.99**	**0.98**	**0.35**	-0.11
$n = 1296$	rpart	0.9	0.97	0.95	0.3	**-0.04**
$p = 7$	ctree	0.9	0.98	0.93	0.31	-0.14
Energy efficiency 2 [210]	GOST	**0.94**	**0.99**	**0.97**	**0.27**	-0.14
$n = 1296$	rpart	0.9	0.97	0.95	0.13	**-0.01**
$p = 7$	ctree	0.9	0.97	0.96	0.21	-0.16
Faceboook Comments [100]	GOST	**0.56**	0.88	0.84	**0.06**	-0.1
$n = 40949$	rpart	**0.56**	0.88	0.84	**0.06**	**-0.09**
$p = 52$	ctree	0.55	**0.89**	**0.86**	**0.06**	-0.11
Faceboook Metrics [139]	GOST	**0.03**	0.55	0.1	**0.01**	**0.05**
$n = 500$	rpart	0.02	**0.56**	**0.14**	**0.01**	**0.05**
$p = 6$	ctree	0.02	0.53	0.05	**0.01**	0.02
Fires	GOST	**0**	**0.5**	**0**	**0**	**0.11**
$n = 517$	rpart	**0**	**0.5**	**0**	**0**	**0.11**
$p = 11$	ctree	**0**	**0.5**	**0**	**0**	**0.11**
GeoMusic [239]	GOST	0.03	0.58	0.32	0.01	0.02
$n = 1059$	rpart	**0.06**	**0.61**	0.37	**0.02**	**0.06**
$p = 115$	ctree	0.03	0.59	**0.38**	0.01	0.03
Insurance Company Benchmark [220]	GOST	0.02	0.59	0.24	**0**	**0.33**
$n = 5822$	rpart	0.02	0.6	0.25	**0**	**0.33**
$p = 84$	ctree	**0.03**	**0.62**	**0.27**	**0**	**0.33**
KEGG Directed [63]	GOST	**0.81**	**0.96**	**0.94**	**0.11**	0.06
$n = 53413$	rpart	0.78	0.95	0.93	**0.11**	**0.07**
$p = 22$	ctree	0.79	0.95	0.93	**0.11**	**0.07**
KEGG Undirected [63]	GOST	**0.87**	0.96	**0.97**	0.14	**0.87**
$n = 65554$	rpart	0.81	0.95	0.94	0.16	0.81
$p = 25$	ctree	0.86	**0.97**	0.96	**0.17**	0.85
Kernel Performance [16]	GOST	**0.73**	**0.84**	**0.84**	**0.09**	**0.53**
$n = 100000$	rpart	0.67	0.8	0.79	0.07	0.43
$p = 13$	ctree	0.69	0.82	0.81	0.08	0.47
Las Vegas Strip [138]	GOST	**0.02**	**0.54**	**0.12**	**0**	**0.02**
$n = 504$	rpart	0	0.51	0.05	**0**	-0.01
$p = 18$	ctree	0.01	0.52	0.05	**0**	0

Continued on next page

Table D.6 – continued from previous page

Dataset	Method	Integrated Brier	Harrell's C	Uno's C	Cox PL	Brier point
Online News Popularity	GOST	**0.05**	0.62	0.56	**0.01**	0.16
$n = 39644$	rpart	**0.05**	0.62	0.57	**0.01**	0.16
$p = 58$	ctree	**0.05**	**0.63**	**0.58**	**0.01**	**0.17**
Online Video Characteristics [63]	GOST	0.75	**0.92**	**0.91**	0.06	**0.76**
$n = 68784$	rpart	0.7	0.91	0.89	**0.15**	0.73
$p = 19$	ctree	**0.76**	**0.92**	**0.91**	**0.15**	**0.76**
Optical Interconnection Network [4]	GOST	-0.08	**0.81**	0.72	**0.12**	**0.7**
$n = 640$	rpart	**0.32**	0.79	0.67	**0.12**	0.68
$p = 8$	ctree	0.27	**0.81**	**0.73**	0.11	0.69
Parkinson Telemonitoring [209]	GOST	**0.59**	**0.85**	**0.8**	**0.14**	**0.67**
$n = 5875$	rpart	0.49	0.82	0.77	0.1	0.55
$p = 18$	ctree	0.42	0.8	0.73	0.08	0.48
PM2.5-Beijing [122]	GOST	**0.35**	**0.8**	**0.77**	0.05	**0.43**
$n = 50387$	rpart	0.32	0.79	0.75	0.05	0.4
$p = 12$	ctree	0.34	**0.8**	0.76	**0.06**	0.42
Propulsion Plant [49]	GOST	**0.64**	**0.88**	**0.87**	**0.11**	**0.7**
$n = 11934$	rpart	0.43	0.8	0.74	0.08	0.46
$p = 15$	ctree	0.28	0.72	0.57	0.04	0.32
Protein [63]	GOST	**0.3**	**0.75**	**0.72**	**0.04**	**0.32**
$n = 45730$	rpart	0.26	0.73	0.69	0.03	0.27
$p = 8$	ctree	0.26	0.72	0.69	0.03	0.27
Real Estate 1 [237]	GOST	**0.44**	**0.83**	**0.67**	**0.18**	**0.59**
$n = 414$	rpart	0.4	0.79	0.58	0.15	0.56
$p = 5$	ctree	0.42	0.8	0.64	0.15	0.56
Real Estate 2 [237]	GOST	**0.8**	0.55	**0.9**	0.02	**0.83**
$n = 53500$	rpart	0.75	**0.92**	0.87	**0.1**	0.76
$p = 383$	ctree	0.76	0.88	0.88	0.05	0.78
Residential Building [182]	GOST	0.63	0.86	0.73	0.24	0.63
$n = 372$	rpart	**0.64**	**0.87**	**0.76**	**0.27**	**0.68**
$p = 107$	ctree	0.62	**0.87**	0.74	0.24	**0.68**
Servo [63]	GOST	**0.51**	**0.85**	**0.75**	**0.26**	**0.48**
$n = 167$	rpart	0.4	0.73	0.5	0.16	0.29
$p = 3$	ctree	0.39	0.69	0.41	0.14	0.24
Stock Market Istanbul [8]	GOST	0.11	0.68	0.47	0.04	**0.18**
$n = 536$	rpart	0.12	**0.69**	**0.5**	**0.05**	0.17
$p = 6$	ctree	**0.14**	**0.69**	**0.5**	**0.05**	0.2
Stock Portfolio [124]	GOST	**0.42**	**0.78**	**0.53**	**0.34**	**0.49**
$n = 63$	rpart	0.26	0.73	0.46	0.26	0.4
$p = 11$	ctree	0.29	0.74	0.43	0.27	0.35
Student Performance [136]	GOST	0.05	0.58	0.2	**0.02**	0.07
$n = 395$	rpart	**0.08**	**0.61**	0.21	**0.02**	**0.11**
$p = 29$	ctree	**0.08**	**0.61**	**0.26**	**0.02**	0.1
Wine Quality [51]	GOST	0.16	0.73	0.63	0.02	**-0.04**

Continued on next page

Table D.6 – continued from previous page

Dataset	Method	Integrated Brier	Harrell's C	Uno's C	Cox PL	Brier point
$n = 6497$	rpart	0.16	0.74	0.65	0.02	**-0.04**
$p = 10$	ctree	**0.18**	**0.76**	**0.71**	**0.03**	-0.07
Yacht [63]	GOST	**0.84**	**0.94**	**0.85**	0.37	0.8
$n = 308$	rpart	0.8	0.91	0.81	**0.41**	0.76
$p = 5$	ctree	0.82	0.9	0.77	0.4	**0.82**

D.3.4 Full Results for UCI Experiments with Simulated Censoring

Table D.7: Dataset specific Brier Score Results for each level of censoring.

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
3D Spatial Network	GOST	0.40	0.42	0.39	0.43	0.43	0.47	0.50	0.48	0.48
	rpart	0.31	0.31	0.31	0.33	0.31	0.34	0.34	0.34	0.37
	ctree	0.36	0.41	0.39	0.39	0.39	0.39	0.41	0.40	0.35
Airfoil Self Noise	GOST	0.47	0.49	0.41	0.40	0.36	0.22	0.35	0.32	0.46
	rpart	0.37	0.35	0.28	0.33	0.22	0.32	0.31	0.33	0.42
	ctree	0.39	0.39	0.37	0.34	0.32	0.29	0.26	0.34	0.46
Appliances Energy Prediction	GOST	0.14	0.14	0.13	0.21	0.22	0.22	0.27	0.24	0.16
	rpart	0.13	0.14	0.16	0.16	0.21	0.22	0.22	0.21	0.17
	ctree	0.15	0.14	0.13	0.17	0.18	0.21	0.23	0.20	0.17
Automobile	GOST	0.00	0.06	0.04	0.00	0.05	0.11	0.00	0.00	0.00
	rpart	0.06	0.19	0.15	0.13	0.05	0.18	0.02	-0.01	-0.10
	ctree	0.08	0.05	0.06	0.07	0.06	0.10	0.08	0.10	-0.10
AutoMPG	GOST	0.59	0.60	0.59	0.54	0.49	0.58	0.49	0.49	0.57
	rpart	0.59	0.57	0.57	0.55	0.45	0.57	0.58	0.60	0.54
	ctree	0.60	0.57	0.54	0.49	0.57	0.54	0.57	0.47	0.58
Behavior Urban Traffic	GOST	0.30	0.26	0.14	0.19	0.25	0.16	0.19	0.11	0.00
	rpart	0.30	0.26	0.14	0.19	0.34	0.23	0.19	0.11	0.07
	ctree	0.30	0.26	0.14	0.25	0.32	0.12	0.18	0.00	0.00
BikeSharing	GOST	0.93	0.90	0.89	0.92	0.91	0.93	0.92	0.94	0.94
	rpart	0.77	0.84	0.87	0.89	0.88	0.92	0.91	0.92	0.94
	ctree	0.94	0.92	0.92	0.93	0.93	0.94	0.94	0.93	0.92
Blog Feedback	GOST	0.29	0.35	0.39	0.40	0.42	0.43	0.41	0.40	0.38
	rpart	0.30	0.35	0.39	0.40	0.42	0.44	0.43	0.41	0.38
	ctree	0.31	0.36	0.37	0.38	0.40	0.41	0.41	0.39	0.37
Buzz in Social Media	GOST	0.74	0.73	0.74	0.76	0.78	0.80	0.80	0.81	0.80
	rpart	0.68	0.68	0.70	0.73	0.77	0.79	0.80	0.81	0.80
	ctree	0.74	0.74	0.73	0.76	0.77	0.79	0.80	0.81	0.78
Cargo2000	GOST	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00
	rpart	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	1.00
	ctree	0.70	0.69	0.79	0.80	0.80	0.82	0.98	1.00	1.00
Communities Crime	GOST	0.68	0.65	0.59	0.61	0.63	0.63	0.64	0.63	0.69
	rpart	0.56	0.63	0.64	0.62	0.69	0.69	0.70	0.63	0.68
	ctree	0.72	0.69	0.67	0.70	0.70	0.68	0.66	0.66	0.72
Computer Hardware	GOST	0.86	0.74	0.84	0.84	0.52	0.80	0.65	0.51	0.48
	rpart	0.72	0.28	0.81	0.61	0.76	0.59	0.60	0.54	0.61
	ctree	0.82	0.85	0.83	0.77	0.52	0.72	0.28	0.74	0.35
Concrete Slump	GOST	0.22	0.11	0.07	0.04	0.00	0.08	0.09	0.00	0.05
	rpart	0.15	0.11	0.07	0.04	0.06	0.08	0.20	0.31	0.18
	ctree	0.00	0.00	0.00	0.04	0.00	0.08	0.09	0.00	0.05
Concrete Strength	GOST	0.52	0.49	0.45	0.31	0.35	0.35	0.45	0.38	0.45
	rpart	0.41	0.39	0.42	0.38	0.37	0.41	0.42	0.40	0.46
	ctree	0.47	0.40	0.41	0.34	0.40	0.45	0.33	0.45	0.37

Continued on next page

Table D.7 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
CSM	GOST	0.26	0.29	0.22	0.30	0.32	0.15	0.18	0.30	0.19
	rpart	0.39	0.33	0.22	0.40	0.35	0.30	0.32	0.30	0.25
	ctree	0.28	0.23	0.37	0.44	0.35	0.16	0.12	0.26	0.02
Cycle Power	GOST	0.76	0.76	0.74	0.73	0.74	0.76	0.75	0.71	0.63
	rpart	0.71	0.67	0.71	0.71	0.75	0.75	0.76	0.69	0.63
	ctree	0.76	0.76	0.75	0.74	0.75	0.76	0.76	0.69	0.61
Electrical Stability	GOST	0.42	0.43	0.41	0.42	0.39	0.40	0.40	0.39	0.34
	rpart	0.35	0.36	0.36	0.35	0.34	0.33	0.35	0.30	0.31
	ctree	0.41	0.41	0.40	0.41	0.41	0.38	0.40	0.37	0.32
Energy Efficiency 1	GOST	0.96	0.95	0.93	0.96	0.93	0.97	0.96	0.93	0.93
	rpart	0.82	0.87	0.89	0.87	0.93	0.94	0.95	0.90	0.89
	ctree	0.94	0.92	0.89	0.89	0.89	0.91	0.92	0.90	0.89
Energy Efficiency 2	GOST	0.94	0.94	0.94	0.90	0.92	0.96	0.96	0.95	0.95
	rpart	0.83	0.84	0.87	0.90	0.90	0.95	0.94	0.95	0.91
	ctree	0.93	0.93	0.89	0.87	0.89	0.89	0.91	0.89	0.88
Faceboook Comments	GOST	0.47	0.53	0.59	0.60	0.60	0.60	0.59	0.57	0.53
	rpart	0.44	0.54	0.59	0.61	0.61	0.59	0.58	0.58	0.52
	ctree	0.44	0.54	0.59	0.60	0.61	0.57	0.55	0.54	0.49
Faceboook Metrics	GOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.14
	rpart	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.11	0.03
	ctree	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14
Fires	GOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	rpart	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ctree	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GeoMusic	GOST	0.02	0.00	0.03	0.00	0.02	0.07	0.07	0.00	0.10
	rpart	0.09	0.02	0.05	0.05	0.06	0.07	0.07	0.02	0.10
	ctree	0.03	0.03	0.04	0.03	0.02	0.03	0.05	0.03	0.04
Insurance Benchmark	GOST	0.02	0.04	0.03	0.00	0.00	0.05	0.00	0.00	0.00
	rpart	0.03	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00
	ctree	0.04	0.04	0.05	0.00	0.00	0.00	0.00	0.05	0.05
KEGG Directed	GOST	0.76	0.76	0.75	0.76	0.77	0.82	0.88	0.92	0.91
	rpart	0.73	0.72	0.73	0.73	0.76	0.78	0.85	0.88	0.88
	ctree	0.73	0.72	0.72	0.73	0.75	0.80	0.87	0.88	0.88
KEGG Undirected	GOST	0.86	0.85	0.85	0.86	0.87	0.87	0.88	0.90	0.91
	rpart	0.75	0.79	0.75	0.81	0.82	0.82	0.85	0.85	0.85
	ctree	0.84	0.84	0.85	0.86	0.87	0.88	0.88	0.88	0.88
Kernel Performance	GOST	0.81	0.81	0.81	0.79	0.77	0.77	0.70	0.61	0.49
	rpart	0.77	0.74	0.77	0.76	0.71	0.71	0.62	0.53	0.39
	ctree	0.79	0.79	0.79	0.79	0.76	0.72	0.66	0.54	0.39
Las Vegas Strip	GOST	0.00	0.00	0.00	0.02	0.08	0.07	0.00	0.00	-0.02
	rpart	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ctree	0.07	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
Online News Popularity	GOST	0.04	0.04	0.04	0.04	0.04	0.05	0.07	0.07	0.08
	rpart	0.04	0.04	0.04	0.04	0.04	0.05	0.06	0.06	0.06
	ctree	0.04	0.04	0.04	0.04	0.05	0.05	0.06	0.07	0.08
Online Video Characteristics	GOST	0.70	0.71	0.75	0.76	0.77	0.76	0.75	0.77	0.79
	rpart	0.62	0.62	0.66	0.71	0.72	0.74	0.73	0.75	0.77
	ctree	0.72	0.72	0.74	0.76	0.77	0.78	0.78	0.77	0.78

Continued on next page

Table D.7 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Optical Network	GOST	0.30	0.55	0.11	-2.01	0.17	-1.17	0.42	0.58	0.33
	rpart	0.01	0.60	0.23	0.34	0.48	0.35	0.20	0.32	0.30
	ctree	0.00	0.10	0.22	0.27	0.45	0.30	0.34	0.37	0.36
Parkinson Telemonitoring	GOST	0.61	0.69	0.55	0.55	0.55	0.56	0.50	0.58	0.69
	rpart	0.48	0.44	0.50	0.53	0.47	0.47	0.42	0.59	0.51
	ctree	0.39	0.39	0.37	0.43	0.45	0.50	0.37	0.45	0.47
PM2.5 - Beijing	GOST	0.32	0.32	0.31	0.32	0.35	0.36	0.39	0.41	0.40
	rpart	0.29	0.28	0.29	0.30	0.32	0.29	0.35	0.39	0.40
	ctree	0.29	0.28	0.29	0.30	0.32	0.35	0.38	0.41	0.40
Propulsion Plant	GOST	0.65	0.66	0.60	0.58	0.65	0.64	0.67	0.67	0.65
	rpart	0.40	0.43	0.44	0.41	0.43	0.35	0.48	0.46	0.45
	ctree	0.34	0.38	0.38	0.31	0.27	0.27	0.27	0.30	0.00
Protein	GOST	0.28	0.28	0.26	0.28	0.31	0.35	0.33	0.33	0.33
	rpart	0.24	0.25	0.25	0.25	0.26	0.29	0.28	0.25	0.23
	ctree	0.25	0.26	0.26	0.27	0.28	0.26	0.27	0.26	0.23
Real Estate 1	GOST	0.45	0.46	0.38	0.31	0.45	0.45	0.54	0.54	0.43
	rpart	0.38	0.29	0.32	0.33	0.36	0.45	0.54	0.46	0.51
	ctree	0.40	0.36	0.42	0.37	0.39	0.43	0.49	0.51	0.43
Real Estate 2	GOST	0.80	0.77	0.77	0.76	0.80	0.79	0.83	0.84	0.83
	rpart	0.69	0.71	0.71	0.70	0.73	0.76	0.81	0.81	0.81
	ctree	0.74	0.74	0.74	0.75	0.75	0.78	0.78	0.79	0.79
ResidentialBuilding	GOST	0.70	0.60	0.60	0.66	0.66	0.72	0.62	0.63	0.50
	rpart	0.52	0.56	0.70	0.75	0.71	0.66	0.74	0.63	0.50
	ctree	0.67	0.69	0.69	0.67	0.69	0.68	0.63	0.49	0.34
Servo	GOST	0.78	0.46	0.76	0.66	0.54	0.54	0.52	0.42	-0.12
	rpart	0.47	0.42	0.48	0.68	0.59	0.41	0.30	0.16	0.09
	ctree	0.76	0.42	0.44	0.52	0.52	0.41	0.30	0.17	0.00
Stockmarket Istanbul	GOST	0.16	0.16	0.08	0.12	0.14	0.05	0.11	0.07	0.10
	rpart	0.10	0.16	0.16	0.22	0.17	0.08	0.11	0.05	0.10
	ctree	0.18	0.15	0.19	0.19	0.15	0.11	0.11	0.10	0.08
Stock Portfolio	GOST	0.73	0.51	0.29	0.48	0.13	0.42	0.00	0.65	0.56
	rpart	0.37	0.00	0.26	0.13	-0.28	0.42	0.22	0.65	0.56
	ctree	0.11	0.21	0.04	0.19	0.13	0.19	0.43	0.60	0.68
Student Performance	GOST	0.07	0.08	0.07	0.04	0.00	0.00	0.06	0.09	0.05
	rpart	0.07	0.08	0.07	0.04	0.08	0.05	0.06	0.09	0.15
	ctree	0.07	0.08	0.07	0.08	0.08	0.06	0.06	0.09	0.12
WineQuality	GOST	0.19	0.17	0.17	0.17	0.17	0.16	0.15	0.16	0.14
	rpart	0.18	0.19	0.18	0.17	0.18	0.16	0.15	0.14	0.13
	ctree	0.21	0.20	0.19	0.18	0.20	0.17	0.16	0.16	0.11
Yacht	GOST	0.91	0.67	0.81	0.82	0.89	0.90	0.91	0.81	0.82
	rpart	0.63	0.67	0.76	0.85	0.89	0.92	0.86	0.80	0.82
	ctree	0.87	0.84	0.84	0.75	0.92	0.90	0.78	0.80	0.70

Table D.8: Dataset specific Harrell's C Score Results for each level of censoring.

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
3D Spatial Network	GOST	0.78	0.79	0.79	0.80	0.80	0.83	0.84	0.85	0.87
	rpart	0.75	0.75	0.76	0.76	0.75	0.77	0.78	0.79	0.82
	ctree	0.75	0.78	0.78	0.79	0.78	0.78	0.80	0.82	0.81
Airfoil Self Noise	GOST	0.81	0.82	0.82	0.81	0.84	0.80	0.83	0.81	0.90
	rpart	0.76	0.76	0.73	0.78	0.74	0.80	0.80	0.81	0.87
	ctree	0.72	0.79	0.79	0.77	0.78	0.80	0.78	0.81	0.80
Appliances Energy Prediction	GOST	0.73	0.73	0.72	0.74	0.74	0.75	0.76	0.78	0.74
	rpart	0.72	0.73	0.72	0.71	0.73	0.74	0.74	0.75	0.77
	ctree	0.72	0.72	0.71	0.73	0.73	0.74	0.76	0.75	0.76
Automobile	GOST	0.50	0.57	0.56	0.50	0.57	0.58	0.50	0.50	0.50
	rpart	0.57	0.71	0.70	0.75	0.57	0.72	0.63	0.65	0.58
	ctree	0.57	0.60	0.57	0.61	0.65	0.66	0.62	0.64	0.58
AutoMPG	GOST	0.87	0.87	0.77	0.76	0.84	0.87	0.86	0.86	0.92
	rpart	0.86	0.85	0.86	0.86	0.81	0.87	0.90	0.89	0.89
	ctree	0.87	0.85	0.86	0.85	0.87	0.85	0.86	0.84	0.93
Behavior Urban Traffic	GOST	0.68	0.69	0.68	0.63	0.63	0.81	0.65	0.67	0.50
	rpart	0.68	0.69	0.68	0.63	0.73	0.67	0.65	0.67	0.66
	ctree	0.68	0.69	0.68	0.68	0.68	0.67	0.67	0.50	0.50
BikeSharing	GOST	0.98	0.98	0.97	0.98	0.97	0.98	0.98	0.99	0.99
	rpart	0.92	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.99
	ctree	0.98	0.97	0.97	0.98	0.98	0.98	0.97	0.98	0.98
Blog Feedback	GOST	0.85	0.84	0.84	0.84	0.84	0.84	0.83	0.84	0.84
	rpart	0.84	0.84	0.84	0.84	0.85	0.85	0.85	0.86	0.85
	ctree	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.86
Buzz in Social Media	GOST	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.93
	rpart	0.91	0.91	0.91	0.91	0.91	0.92	0.92	0.92	0.92
	ctree	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
Cargo2000	GOST	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	rpart	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	ctree	0.89	0.88	0.93	0.93	0.93	0.96	1.00	1.00	1.00
Communities Crime	GOST	0.88	0.89	0.87	0.86	0.87	0.88	0.90	0.91	0.92
	rpart	0.81	0.89	0.89	0.87	0.90	0.90	0.91	0.92	0.95
	ctree	0.92	0.90	0.90	0.91	0.91	0.91	0.90	0.93	0.94
Computer Hardware	GOST	0.92	0.91	0.88	0.88	0.80	0.90	0.85	0.83	0.83
	rpart	0.85	0.70	0.86	0.82	0.89	0.78	0.86	0.81	0.90
	ctree	0.90	0.90	0.90	0.80	0.81	0.81	0.88	0.90	0.75
Concrete Slump	GOST	0.64	0.64	0.64	0.61	0.50	0.65	0.71	0.50	0.66
	rpart	0.62	0.64	0.64	0.61	0.63	0.65	0.67	0.77	0.72
	ctree	0.50	0.50	0.50	0.61	0.50	0.65	0.66	0.50	0.66
Concrete Strength	GOST	0.84	0.85	0.85	0.79	0.80	0.82	0.86	0.84	0.89
	rpart	0.78	0.80	0.82	0.81	0.80	0.83	0.85	0.87	0.88
	ctree	0.83	0.77	0.83	0.81	0.84	0.85	0.81	0.85	0.82
CSM	GOST	0.74	0.74	0.70	0.71	0.72	0.67	0.69	0.72	0.73
	rpart	0.77	0.76	0.70	0.79	0.78	0.73	0.75	0.72	0.82
	ctree	0.72	0.70	0.74	0.75	0.75	0.67	0.74	0.77	0.76
Cycle Power	GOST	0.91	0.91	0.91	0.91	0.92	0.92	0.90	0.93	0.92
	rpart	0.89	0.87	0.90	0.90	0.92	0.92	0.92	0.92	0.93

Continued on next page

Table D.8 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	ctree	0.91	0.92	0.91	0.92	0.92	0.92	0.92	0.93	0.93
Electrical Stability	GOST	0.80	0.81	0.81	0.82	0.82	0.82	0.83	0.83	0.83
	rpart	0.76	0.78	0.79	0.78	0.78	0.79	0.80	0.78	0.82
	ctree	0.80	0.80	0.80	0.82	0.82	0.83	0.84	0.84	0.85
Energy Efficiency 1	GOST	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.98	0.99
	rpart	0.91	0.96	0.96	0.97	0.98	0.98	0.99	0.98	0.97
	ctree	0.99	0.98	0.98	0.97	0.97	0.97	0.98	0.98	0.97
Energy Efficiency 2	GOST	0.99	0.99	0.98	0.96	0.99	0.99	0.99	0.99	0.99
	rpart	0.94	0.94	0.95	0.98	0.97	0.98	0.98	0.98	0.97
	ctree	0.98	0.98	0.97	0.96	0.96	0.96	0.96	0.96	0.97
Faceboook Comments	GOST	0.88	0.88	0.87	0.88	0.88	0.88	0.88	0.89	0.89
	rpart	0.88	0.87	0.88	0.88	0.88	0.88	0.88	0.89	0.89
	ctree	0.88	0.88	0.88	0.89	0.89	0.89	0.89	0.89	0.89
Faceboook Metrics	GOST	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.70	0.74
	rpart	0.50	0.50	0.50	0.62	0.50	0.50	0.50	0.70	0.76
	ctree	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.74
Fires	GOST	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	rpart	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	ctree	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
GeoMusic	GOST	0.60	0.56	0.56	0.55	0.58	0.62	0.62	0.50	0.65
	rpart	0.65	0.57	0.60	0.58	0.61	0.62	0.62	0.57	0.65
	ctree	0.59	0.59	0.60	0.59	0.58	0.58	0.62	0.58	0.58
Insurance Benchmark	GOST	0.72	0.73	0.66	0.50	0.50	0.72	0.50	0.50	0.50
	rpart	0.73	0.73	0.73	0.70	0.50	0.50	0.50	0.50	0.50
	ctree	0.70	0.70	0.71	0.50	0.50	0.50	0.50	0.73	0.71
KEGG Directed	GOST	0.92	0.93	0.94	0.95	0.96	0.97	0.99	0.99	0.99
	rpart	0.91	0.92	0.93	0.94	0.95	0.96	0.98	0.99	0.99
	ctree	0.91	0.92	0.93	0.94	0.95	0.97	0.99	0.99	0.99
KEGG Undirected	GOST	0.94	0.94	0.94	0.97	0.97	0.97	0.98	0.98	0.99
	rpart	0.93	0.95	0.92	0.95	0.96	0.96	0.97	0.98	0.98
	ctree	0.95	0.95	0.96	0.96	0.97	0.97	0.98	0.98	0.98
Kernel Performance	GOST	0.84	0.84	0.83	0.84	0.84	0.84	0.84	0.85	0.85
	rpart	0.81	0.79	0.81	0.81	0.80	0.81	0.80	0.81	0.81
	ctree	0.82	0.82	0.82	0.83	0.83	0.82	0.81	0.82	0.81
Las Vegas Strip	GOST	0.50	0.50	0.50	0.53	0.65	0.64	0.50	0.50	0.58
	rpart	0.61	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	ctree	0.64	0.50	0.50	0.53	0.50	0.50	0.50	0.50	0.50
Online News Popularity	GOST	0.61	0.61	0.61	0.61	0.61	0.61	0.62	0.62	0.63
	rpart	0.61	0.61	0.61	0.61	0.61	0.61	0.62	0.62	0.63
	ctree	0.62	0.62	0.62	0.63	0.63	0.63	0.63	0.63	0.64
Online Video Characteristics	GOST	0.92	0.91	0.92	0.92	0.92	0.91	0.88	0.95	0.95
	rpart	0.90	0.88	0.88	0.90	0.91	0.92	0.93	0.94	0.96
	ctree	0.89	0.91	0.91	0.91	0.92	0.93	0.94	0.95	0.96
Optical Network	GOST	0.88	0.81	0.80	0.80	0.68	0.85	0.79	0.87	0.77
	rpart	0.75	0.80	0.86	0.79	0.79	0.81	0.81	0.81	0.72
	ctree	0.80	0.84	0.84	0.83	0.80	0.81	0.80	0.80	0.77
Parkinson Telemonitoring	GOST	0.80	0.88	0.81	0.83	0.82	0.85	0.85	0.89	0.94
	rpart	0.77	0.77	0.79	0.82	0.81	0.82	0.83	0.89	0.91

Continued on next page

Table D.8 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	ctree	0.75	0.75	0.76	0.78	0.80	0.83	0.81	0.83	0.88
PM2.5 - Beijing	GOST	0.76	0.77	0.78	0.78	0.79	0.80	0.82	0.84	0.86
	rpart	0.75	0.76	0.76	0.77	0.78	0.76	0.80	0.83	0.85
	ctree	0.75	0.76	0.77	0.78	0.79	0.80	0.81	0.83	0.86
Propulsion Plant	GOST	0.87	0.88	0.87	0.88	0.86	0.87	0.92	0.85	0.95
	rpart	0.75	0.77	0.79	0.78	0.80	0.74	0.84	0.85	0.88
	ctree	0.73	0.75	0.77	0.74	0.71	0.73	0.74	0.80	0.50
Protein	GOST	0.72	0.72	0.72	0.74	0.75	0.77	0.76	0.78	0.81
	rpart	0.69	0.71	0.71	0.72	0.72	0.74	0.74	0.74	0.76
	ctree	0.70	0.71	0.71	0.72	0.72	0.72	0.73	0.75	0.75
Real Estate 1	GOST	0.79	0.80	0.80	0.82	0.83	0.84	0.87	0.86	0.84
	rpart	0.79	0.71	0.72	0.73	0.76	0.84	0.87	0.81	0.89
	ctree	0.78	0.77	0.79	0.78	0.81	0.83	0.82	0.82	0.86
Real Estate 2	GOST	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.95	0.50
	rpart	0.88	0.89	0.90	0.90	0.90	0.92	0.94	0.95	0.97
	ctree	0.50	0.88	0.90	0.92	0.92	0.93	0.93	0.94	0.96
ResidentialBuilding	GOST	0.89	0.88	0.84	0.86	0.86	0.88	0.82	0.87	0.88
	rpart	0.75	0.85	0.91	0.92	0.90	0.86	0.92	0.87	0.88
	ctree	0.88	0.88	0.88	0.87	0.85	0.90	0.85	0.85	0.85
Servo	GOST	0.88	0.82	0.88	0.72	0.90	0.85	0.89	0.87	0.85
	rpart	0.69	0.70	0.70	0.80	0.82	0.70	0.70	0.79	0.66
	ctree	0.81	0.70	0.70	0.71	0.71	0.70	0.70	0.69	0.50
Stockmarket Istanbul	GOST	0.70	0.72	0.64	0.69	0.72	0.66	0.70	0.63	0.65
	rpart	0.62	0.71	0.71	0.73	0.72	0.65	0.66	0.71	0.65
	ctree	0.72	0.71	0.70	0.70	0.71	0.69	0.70	0.71	0.61
Stock Portfolio	GOST	0.88	0.81	0.77	0.84	0.74	0.70	0.50	0.88	0.88
	rpart	0.74	0.50	0.75	0.57	0.79	0.70	0.73	0.88	0.88
	ctree	0.59	0.70	0.58	0.75	0.74	0.75	0.78	0.88	0.91
Student Performance	GOST	0.59	0.61	0.60	0.55	0.50	0.50	0.58	0.64	0.68
	rpart	0.59	0.61	0.60	0.55	0.61	0.56	0.58	0.64	0.73
	ctree	0.59	0.61	0.60	0.61	0.61	0.61	0.58	0.64	0.68
WineQuality	GOST	0.75	0.72	0.73	0.75	0.73	0.72	0.71	0.73	0.70
	rpart	0.75	0.76	0.75	0.74	0.74	0.73	0.73	0.72	0.71
	ctree	0.78	0.77	0.76	0.76	0.76	0.76	0.75	0.75	0.74
Yacht	GOST	0.96	0.91	0.95	0.93	0.94	0.91	0.95	0.94	0.95
	rpart	0.87	0.86	0.88	0.93	0.94	0.95	0.91	0.91	0.95
	ctree	0.94	0.91	0.89	0.92	0.89	0.89	0.83	0.92	0.89

Table D.9: Dataset specific Uno's C Score Results for each level of censoring.

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
3D Spatial Network	GOST	0.77	0.78	0.76	0.77	0.77	0.80	0.81	0.82	0.84
	rpart	0.72	0.72	0.72	0.72	0.70	0.72	0.73	0.74	0.78
	ctree	0.73	0.77	0.76	0.76	0.75	0.74	0.77	0.78	0.76

Continued on next page

Table D.9 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Airfoil Self Noise	GOST	0.81	0.82	0.80	0.80	0.78	0.77	0.72	0.66	0.79
	rpart	0.71	0.70	0.67	0.71	0.66	0.70	0.70	0.65	0.78
	ctree	0.71	0.75	0.74	0.71	0.67	0.69	0.66	0.68	0.81
Appliances Energy Prediction	GOST	0.71	0.70	0.69	0.71	0.71	0.70	0.72	0.74	0.65
	rpart	0.70	0.70	0.69	0.64	0.69	0.69	0.69	0.70	0.70
	ctree	0.71	0.70	0.68	0.69	0.69	0.70	0.71	0.71	0.70
Automobile	GOST	0.00	0.18	0.17	0.00	0.18	0.17	0.00	0.00	0.00
	rpart	0.17	0.62	0.63	0.62	0.18	0.54	0.35	0.36	0.18
	ctree	0.18	0.31	0.18	0.25	0.38	0.37	0.26	0.34	0.18
AutoMPG	GOST	0.83	0.82	0.83	0.73	0.72	0.77	0.77	0.75	0.85
	rpart	0.78	0.77	0.77	0.77	0.73	0.77	0.80	0.80	0.82
	ctree	0.82	0.81	0.77	0.77	0.78	0.74	0.77	0.62	0.87
Behavior Urban Traffic	GOST	0.40	0.42	0.43	0.31	0.29	0.63	0.43	0.41	0.00
	rpart	0.40	0.42	0.43	0.31	0.46	0.43	0.43	0.41	0.40
	ctree	0.40	0.42	0.43	0.40	0.41	0.43	0.47	0.00	0.00
BikeSharing	GOST	0.98	0.96	0.95	0.96	0.95	0.96	0.95	0.97	0.99
	rpart	0.85	0.92	0.93	0.93	0.92	0.95	0.95	0.96	0.98
	ctree	0.97	0.97	0.96	0.96	0.96	0.96	0.95	0.96	0.94
Blog Feedback	GOST	0.83	0.80	0.80	0.81	0.79	0.78	0.77	0.78	0.78
	rpart	0.82	0.80	0.80	0.80	0.81	0.81	0.79	0.81	0.80
	ctree	0.83	0.83	0.83	0.82	0.82	0.82	0.82	0.81	0.83
Buzz in Social Media	GOST	0.92	0.91	0.91	0.90	0.90	0.90	0.90	0.90	0.91
	rpart	0.88	0.88	0.87	0.88	0.88	0.89	0.89	0.89	0.90
	ctree	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.89	0.88
Cargo2000	GOST	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	rpart	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
	ctree	0.77	0.77	0.88	0.88	0.88	0.94	0.95	1.00	1.00
Communities Crime	GOST	0.88	0.84	0.77	0.78	0.78	0.77	0.79	0.78	0.88
	rpart	0.79	0.83	0.84	0.78	0.84	0.84	0.85	0.79	0.92
	ctree	0.89	0.87	0.84	0.86	0.86	0.84	0.80	0.83	0.87
Computer Hardware	GOST	0.85	0.82	0.74	0.76	0.64	0.79	0.73	0.69	0.63
	rpart	0.73	0.41	0.73	0.62	0.77	0.54	0.78	0.65	0.77
	ctree	0.83	0.82	0.81	0.66	0.68	0.61	0.72	0.77	0.39
Concrete Slump	GOST	0.35	0.34	0.35	0.26	0.00	0.31	0.46	0.00	0.33
	rpart	0.33	0.34	0.35	0.26	0.27	0.31	0.37	0.58	0.37
	ctree	0.00	0.00	0.00	0.26	0.00	0.31	0.35	0.00	0.33
Concrete Strength	GOST	0.84	0.82	0.80	0.67	0.68	0.69	0.76	0.69	0.73
	rpart	0.74	0.73	0.74	0.73	0.74	0.75	0.76	0.74	0.72
	ctree	0.81	0.78	0.79	0.75	0.78	0.77	0.77	0.72	0.58
CSM	GOST	0.57	0.60	0.45	0.44	0.45	0.47	0.42	0.44	0.44
	rpart	0.64	0.62	0.45	0.65	0.58	0.46	0.54	0.44	0.70
	ctree	0.54	0.53	0.63	0.55	0.59	0.43	0.58	0.66	0.62
Cycle Power	GOST	0.91	0.90	0.88	0.86	0.87	0.87	0.89	0.91	0.89
	rpart	0.83	0.83	0.83	0.83	0.88	0.87	0.88	0.90	0.90
	ctree	0.90	0.89	0.89	0.88	0.88	0.89	0.88	0.90	0.90
Electrical Stability	GOST	0.79	0.79	0.78	0.79	0.78	0.78	0.79	0.79	0.79
	rpart	0.74	0.75	0.75	0.75	0.74	0.74	0.75	0.75	0.78
	ctree	0.79	0.79	0.79	0.79	0.79	0.78	0.79	0.79	0.80

Continued on next page

Table D.9 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Energy Efficiency 1	GOST	0.99	0.97	0.97	0.98	0.97	0.98	0.97	0.97	0.98
	rpart	0.89	0.93	0.94	0.95	0.97	0.97	0.98	0.97	0.97
	ctree	0.98	0.97	0.96	0.95	0.96	0.95	0.97	0.97	0.64
Energy Efficiency 2	GOST	0.98	0.98	0.96	0.95	0.97	0.97	0.99	0.99	0.98
	rpart	0.92	0.90	0.93	0.95	0.96	0.98	0.98	0.98	0.97
	ctree	0.98	0.97	0.95	0.95	0.95	0.94	0.96	0.97	0.96
Faceboook Comments	GOST	0.87	0.84	0.83	0.84	0.83	0.83	0.82	0.84	0.84
	rpart	0.84	0.83	0.83	0.84	0.83	0.83	0.82	0.85	0.84
	ctree	0.86	0.86	0.86	0.86	0.87	0.87	0.86	0.86	0.86
Faceboook Metrics	GOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.41	0.47
	rpart	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.41	0.49
	ctree	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.47
Fires	GOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	rpart	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ctree	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GeoMusic	GOST	0.44	0.31	0.33	0.32	0.38	0.35	0.35	0.00	0.37
	rpart	0.54	0.32	0.43	0.28	0.35	0.35	0.35	0.31	0.37
	ctree	0.46	0.42	0.42	0.38	0.36	0.33	0.36	0.31	0.35
Insurance Benchmark	GOST	0.57	0.58	0.50	0.00	0.00	0.51	0.00	0.00	0.00
	rpart	0.57	0.58	0.58	0.48	0.00	0.00	0.00	0.00	0.00
	ctree	0.48	0.49	0.49	0.00	0.00	0.00	0.00	0.52	0.48
KEGG Directed	GOST	0.92	0.92	0.92	0.92	0.92	0.94	0.98	0.99	0.99
	rpart	0.90	0.90	0.90	0.90	0.91	0.92	0.97	0.99	0.99
	ctree	0.90	0.90	0.90	0.90	0.91	0.94	0.97	0.98	0.98
KEGG Undirected	GOST	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.98	0.99
	rpart	0.92	0.93	0.89	0.93	0.93	0.94	0.95	0.96	0.97
	ctree	0.94	0.94	0.95	0.95	0.96	0.96	0.96	0.97	0.97
Kernel Performance	GOST	0.84	0.83	0.83	0.83	0.84	0.84	0.84	0.84	0.83
	rpart	0.80	0.78	0.80	0.80	0.80	0.79	0.79	0.80	0.78
	ctree	0.81	0.81	0.81	0.82	0.82	0.81	0.80	0.81	0.79
Las Vegas Strip	GOST	0.00	0.00	0.00	0.09	0.44	0.43	0.00	0.00	0.13
	rpart	0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ctree	0.41	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00
Online News Popularity	GOST	0.56	0.56	0.54	0.56	0.54	0.54	0.56	0.56	0.58
	rpart	0.57	0.58	0.57	0.57	0.57	0.57	0.59	0.56	0.53
	ctree	0.59	0.58	0.58	0.59	0.59	0.57	0.58	0.56	0.57
Online Video Characteristics	GOST	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.93	0.94
	rpart	0.88	0.87	0.87	0.88	0.88	0.90	0.90	0.92	0.94
	ctree	0.87	0.89	0.90	0.90	0.90	0.91	0.92	0.92	0.94
Optical Network	GOST	0.87	0.76	0.70	0.74	0.47	0.81	0.69	0.81	0.60
	rpart	0.58	0.73	0.81	0.68	0.69	0.74	0.68	0.66	0.41
	ctree	0.71	0.80	0.78	0.77	0.74	0.73	0.70	0.69	0.67
Parkinson Telemonitoring	GOST	0.81	0.87	0.79	0.79	0.80	0.77	0.76	0.75	0.89
	rpart	0.75	0.74	0.77	0.77	0.75	0.77	0.74	0.83	0.85
	ctree	0.73	0.73	0.73	0.75	0.76	0.79	0.67	0.67	0.77
PM2.5 - Beijing	GOST	0.75	0.76	0.75	0.75	0.76	0.76	0.77	0.80	0.83
	rpart	0.73	0.73	0.73	0.73	0.73	0.71	0.75	0.78	0.82
	ctree	0.74	0.74	0.74	0.75	0.75	0.76	0.77	0.79	0.82

Continued on next page

Table D.9 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Propulsion Plant	GOST	0.87	0.87	0.84	0.84	0.87	0.87	0.89	0.90	0.91
	rpart	0.70	0.72	0.73	0.71	0.73	0.68	0.76	0.80	0.80
	ctree	0.64	0.70	0.70	0.64	0.56	0.59	0.60	0.73	0.00
Protein	GOST	0.69	0.71	0.68	0.70	0.72	0.74	0.74	0.75	0.78
	rpart	0.67	0.68	0.68	0.68	0.68	0.71	0.71	0.70	0.69
	ctree	0.68	0.69	0.68	0.69	0.69	0.69	0.70	0.71	0.70
Real Estate 1	GOST	0.70	0.66	0.64	0.71	0.65	0.66	0.68	0.64	0.72
	rpart	0.72	0.45	0.45	0.45	0.46	0.66	0.68	0.58	0.77
	ctree	0.65	0.60	0.67	0.59	0.63	0.67	0.65	0.56	0.75
Real Estate 2	GOST	0.91	0.89	0.89	0.88	0.89	0.90	0.92	0.93	0.94
	rpart	0.82	0.83	0.83	0.83	0.84	0.88	0.90	0.92	0.93
	ctree	0.86	0.86	0.88	0.88	0.87	0.89	0.89	0.90	0.91
ResidentialBuilding	GOST	0.90	0.75	0.67	0.69	0.68	0.76	0.62	0.72	0.80
	rpart	0.69	0.69	0.82	0.82	0.79	0.72	0.84	0.72	0.80
	ctree	0.78	0.77	0.75	0.73	0.68	0.78	0.73	0.70	0.73
Servo	GOST	0.83	0.71	0.83	0.46	0.85	0.74	0.82	0.74	0.76
	rpart	0.40	0.40	0.41	0.76	0.77	0.44	0.43	0.55	0.33
	ctree	0.73	0.40	0.40	0.43	0.43	0.44	0.43	0.42	0.00
Stockmarket Istanbul	GOST	0.58	0.60	0.40	0.51	0.53	0.46	0.55	0.26	0.31
	rpart	0.46	0.59	0.61	0.63	0.57	0.42	0.41	0.50	0.31
	ctree	0.61	0.57	0.56	0.55	0.57	0.52	0.48	0.49	0.18
Stock Portfolio	GOST	0.75	0.62	0.61	0.64	0.40	0.31	0.00	0.74	0.68
	rpart	0.49	0.00	0.48	0.38	0.60	0.31	0.49	0.74	0.68
	ctree	0.18	0.42	0.16	0.51	0.40	0.44	0.41	0.53	0.85
Student Performance	GOST	0.26	0.27	0.26	0.11	0.00	0.00	0.11	0.28	0.51
	rpart	0.26	0.27	0.26	0.11	0.26	0.10	0.11	0.28	0.26
	ctree	0.26	0.27	0.26	0.26	0.26	0.25	0.11	0.28	0.41
WineQuality	GOST	0.69	0.60	0.61	0.70	0.61	0.60	0.60	0.64	0.61
	rpart	0.67	0.70	0.69	0.67	0.66	0.63	0.63	0.62	0.60
	ctree	0.73	0.73	0.71	0.71	0.70	0.71	0.70	0.70	0.71
Yacht	GOST	0.93	0.81	0.90	0.84	0.87	0.83	0.91	0.81	0.77
	rpart	0.73	0.76	0.80	0.84	0.87	0.90	0.83	0.78	0.77
	ctree	0.89	0.82	0.77	0.81	0.80	0.79	0.66	0.75	0.69

Table D.10: Dataset specific Cox Score Results for each level of censoring.

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
3D Spatial Network	GOST	0.04	0.04	0.03	0.06	0.02	0.06	0.08	0.06	0.09
	rpart	0.03	0.04	0.04	0.04	0.04	0.05	0.05	0.06	0.08
	ctree	0.04	0.05	0.05	0.05	0.05	0.05	0.06	0.07	0.07
Airfoil Self Noise	GOST	0.05	0.05	0.07	0.05	0.14	0.06	0.13	0.12	0.11
	rpart	0.07	0.07	0.06	0.08	0.06	0.10	0.10	0.11	0.18
	ctree	0.05	0.09	0.09	0.09	0.09	0.10	0.08	0.11	0.03
Appliances Energy Prediction	GOST	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.05	0.04
	rpart	0.02	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.05

Continued on next page

Table D.10 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	ctree	0.02	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.04
Automobile	GOST	0.00	0.02	0.02	0.00	0.02	0.02	0.00	0.00	0.00
	rpart	0.02	0.07	0.05	0.09	0.02	0.07	0.04	0.05	0.05
	ctree	0.01	0.01	0.01	0.02	0.03	0.04	0.08	0.05	0.05
AutoMPG	GOST	0.19	0.19	0.12	0.04	0.16	0.22	0.22	0.21	0.34
	rpart	0.15	0.15	0.15	0.16	0.14	0.21	0.27	0.28	0.31
	ctree	0.21	0.19	0.19	0.18	0.23	0.18	0.20	0.19	0.35
Behavior Urban Traffic	GOST	0.11	0.11	0.08	0.08	0.06	0.17	0.08	0.07	0.00
	rpart	0.11	0.11	0.08	0.08	0.13	0.13	0.08	0.07	0.05
	ctree	0.11	0.11	0.08	0.11	0.12	0.08	0.08	0.00	0.00
BikeSharing	GOST	0.19	0.04	0.25	0.10	0.09	0.30	0.26	0.03	0.11
	rpart	0.18	0.04	0.04	0.03	0.10	0.04	0.11	0.12	0.17
	ctree	0.24	0.19	0.17	0.03	0.09	0.19	0.18	0.34	0.36
Blog Feedback	GOST	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.04
	rpart	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.04	0.04
	ctree	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.04
Buzz in Social Media	GOST	0.11	0.11	0.11	0.12	0.12	0.13	0.14	0.14	0.15
	rpart	0.10	0.10	0.11	0.11	0.12	0.12	0.13	0.14	0.15
	ctree	0.11	0.11	0.11	0.12	0.12	0.13	0.13	0.14	0.14
Cargo2000	GOST	0.15	0.16	0.18	0.19	0.21	0.23	0.24	0.26	0.28
	rpart	0.15	0.16	0.18	0.20	0.21	0.23	0.22	0.26	0.28
	ctree	0.08	0.09	0.10	0.12	0.13	0.16	0.24	0.26	0.28
Communities Crime	GOST	0.15	0.14	0.14	0.15	0.16	0.16	0.17	0.19	0.27
	rpart	0.10	0.14	0.15	0.13	0.16	0.17	0.20	0.21	0.29
	ctree	0.19	0.18	0.17	0.16	0.18	0.19	0.19	0.22	0.29
Computer Hardware	GOST	0.19	0.18	0.17	0.34	0.23	0.42	0.28	0.07	0.30
	rpart	0.29	0.06	0.30	0.22	0.34	0.24	0.30	0.27	0.38
	ctree	0.34	0.40	0.36	0.22	0.01	0.29	0.37	0.43	0.23
Concrete Slump	GOST	0.05	0.04	0.04	0.03	0.00	0.06	0.07	0.00	0.05
	rpart	0.04	0.04	0.04	0.03	0.04	0.06	0.08	0.15	0.12
	ctree	0.00	0.00	0.00	0.03	0.00	0.06	0.06	0.00	0.05
Concrete Strength	GOST	0.12	0.03	0.03	0.09	0.10	0.12	0.16	0.14	0.21
	rpart	0.10	0.10	0.11	0.11	0.11	0.14	0.13	0.16	0.20
	ctree	0.13	0.05	0.13	0.12	0.12	0.14	0.05	0.15	0.16
CSM	GOST	0.08	0.09	0.07	0.08	0.09	0.05	0.06	0.11	0.12
	rpart	0.10	0.09	0.07	0.12	0.11	0.09	0.11	0.11	0.18
	ctree	0.06	0.06	0.09	0.11	0.10	0.05	0.07	0.15	0.13
Cycle Power	GOST	0.13	0.13	0.17	0.17	0.18	0.19	0.12	0.21	0.13
	rpart	0.14	0.13	0.15	0.16	0.18	0.19	0.20	0.20	0.21
	ctree	0.16	0.17	0.17	0.17	0.18	0.19	0.20	0.20	0.21
Electrical Stability	GOST	0.07	0.07	0.07	0.08	0.08	0.08	0.09	0.10	0.10
	rpart	0.05	0.06	0.06	0.06	0.06	0.06	0.08	0.07	0.09
	ctree	0.06	0.07	0.07	0.08	0.08	0.08	0.09	0.09	0.11
Energy Efficiency 1	GOST	0.31	0.21	0.27	0.32	0.30	0.31	0.37	0.41	0.64
	rpart	0.18	0.21	0.23	0.24	0.27	0.29	0.34	0.39	0.56
	ctree	0.28	0.25	0.24	0.24	0.25	0.27	0.33	0.40	0.56
Energy Efficiency 2	GOST	0.14	0.07	0.25	0.11	0.27	0.16	0.32	0.42	0.64
	rpart	0.06	0.01	0.16	0.05	0.00	0.12	0.10	0.16	0.53

Continued on next page

Table D.10 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	ctree	0.14	0.09	0.00	0.17	0.22	0.19	0.26	0.34	0.49
Faceboook Comments	GOST	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
	rpart	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
	ctree	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Faceboook Metrics	GOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.08
	rpart	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.05	0.08
	ctree	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
Fires	GOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	rpart	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ctree	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GeoMusic	GOST	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.00	0.03
	rpart	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.03
	ctree	0.01	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02
Insurance Benchmark	GOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	rpart	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ctree	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KEGG Directed	GOST	0.09	0.10	0.10	0.10	0.11	0.12	0.13	0.14	0.14
	rpart	0.09	0.09	0.09	0.10	0.11	0.10	0.12	0.13	0.14
	ctree	0.09	0.09	0.09	0.10	0.10	0.11	0.13	0.13	0.13
KEGG Undirected	GOST	0.05	0.05	0.04	0.17	0.18	0.18	0.20	0.21	0.22
	rpart	0.13	0.14	0.13	0.15	0.16	0.17	0.18	0.19	0.21
	ctree	0.15	0.15	0.16	0.16	0.17	0.18	0.19	0.20	0.21
Kernel Performance	GOST	0.09	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.09
	rpart	0.07	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.07
	ctree	0.07	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.07
Las Vegas Strip	GOST	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00
	rpart	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ctree	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Online News Popularity	GOST	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	rpart	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	ctree	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Online Video Characteristics	GOST	0.05	0.14	0.15	0.15	0.16	0.12	0.08	0.02	-0.35
	rpart	0.12	0.11	0.10	0.13	0.14	0.15	0.17	0.19	0.22
	ctree	0.11	0.13	0.13	0.14	0.15	0.16	0.17	0.19	0.21
Optical Network	GOST	0.09	0.15	0.11	0.15	0.07	0.10	0.10	0.25	0.06
	rpart	0.05	0.11	0.16	0.12	0.12	0.12	0.12	0.14	0.10
	ctree	0.08	0.14	0.12	0.10	0.10	0.11	0.11	0.14	0.12
Parkinson Telemonitoring	GOST	0.07	0.15	0.11	0.11	0.07	0.14	0.14	0.18	0.26
	rpart	0.07	0.06	0.08	0.10	0.09	0.10	0.10	0.17	0.18
	ctree	0.06	0.05	0.05	0.07	0.08	0.10	0.09	0.12	0.15
PM2.5 - Beijing	GOST	0.04	0.04	0.04	0.05	0.05	0.06	0.07	0.08	0.02
	rpart	0.04	0.04	0.04	0.04	0.05	0.04	0.06	0.07	0.09
	ctree	0.03	0.04	0.04	0.05	0.05	0.06	0.06	0.08	0.09
Propulsion Plant	GOST	0.05	0.06	0.11	0.11	0.11	0.08	0.18	0.10	0.23
	rpart	0.05	0.06	0.06	0.06	0.08	0.06	0.11	0.12	0.14
	ctree	0.03	0.05	0.06	0.04	0.04	0.05	0.05	0.08	0.00
Protein	GOST	0.03	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.07
	rpart	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.04	0.05

Continued on next page

Table D.10 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
	ctree	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.05
Real Estate 1	GOST	0.13	0.14	0.12	0.16	0.16	0.18	0.22	0.25	0.24
	rpart	0.13	0.05	0.08	0.09	0.11	0.18	0.22	0.19	0.32
	ctree	0.10	0.07	0.12	0.10	0.14	0.18	0.17	0.20	0.25
Real Estate 2	GOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00
	rpart	0.04	0.12	0.04	0.00	0.09	0.05	0.05	0.24	0.28
	ctree	0.00	0.11	0.04	0.05	0.05	0.04	0.05	0.06	0.06
ResidentialBuilding	GOST	0.09	0.23	0.20	0.23	0.22	0.28	0.23	0.30	0.33
	rpart	0.14	0.22	0.29	0.29	0.27	0.24	0.32	0.30	0.33
	ctree	0.23	0.23	0.25	0.24	0.24	0.28	0.25	0.21	0.21
Servo	GOST	0.31	0.21	0.31	0.21	0.21	0.27	0.34	0.30	0.21
	rpart	0.13	0.13	0.13	0.27	0.21	0.14	0.15	0.16	0.11
	ctree	0.24	0.13	0.13	0.16	0.17	0.14	0.15	0.14	0.00
Stockmarket Istanbul	GOST	0.05	0.05	0.03	0.04	0.05	0.04	0.06	0.03	0.05
	rpart	0.03	0.05	0.05	0.06	0.05	0.03	0.04	0.05	0.05
	ctree	0.06	0.05	0.06	0.06	0.05	0.05	0.06	0.05	0.03
Stock Portfolio	GOST	0.48	0.35	0.34	0.40	0.23	0.24	0.00	0.55	0.47
	rpart	0.24	0.00	0.25	0.06	0.24	0.24	0.26	0.55	0.47
	ctree	0.10	0.21	0.09	0.26	0.23	0.26	0.25	0.49	0.57
Student Performance	GOST	0.01	0.02	0.01	0.01	0.00	0.00	0.02	0.03	0.06
	rpart	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.03	0.03
	ctree	0.01	0.02	0.01	0.01	0.01	0.02	0.02	0.03	0.06
WineQuality	GOST	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03
	rpart	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.04
	ctree	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04
Yacht	GOST	0.56	0.23	0.28	0.24	0.46	0.42	0.31	0.31	0.51
	rpart	0.31	0.31	0.33	0.45	0.46	0.50	0.42	0.40	0.51
	ctree	0.49	0.42	0.38	0.44	0.40	0.39	0.29	0.42	0.37

Table D.11: Dataset specific Brier Point Ratio Results for each level of censoring.

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
3D Spatial Network	GOST	0.44	0.46	0.45	0.47	0.48	0.53	0.51	0.49	0.45
	rpart	0.36	0.36	0.37	0.39	0.33	0.35	0.34	0.35	0.33
	ctree	0.38	0.46	0.43	0.46	0.42	0.39	0.41	0.41	0.33
Airfoil Self Noise	GOST	0.52	0.52	0.54	0.61	0.58	0.53	0.48	0.46	0.52
	rpart	0.37	0.41	0.38	0.52	0.33	0.47	0.47	0.44	0.41
	ctree	0.41	0.46	0.52	0.49	0.45	0.47	0.44	0.42	0.51
Appliances Energy Prediction	GOST	0.20	0.21	0.16	0.18	0.23	0.05	0.06	0.22	-0.06
	rpart	0.18	0.20	0.18	0.18	0.24	0.02	0.03	0.19	-0.04
	ctree	0.18	0.16	0.16	0.16	0.22	0.04	0.04	0.17	-0.07
Automobile	GOST	0.00	0.02	-0.06	0.00	-0.05	0.07	0.00	0.00	0.00
	rpart	0.08	0.29	0.33	0.19	-0.05	0.14	0.10	-0.04	-0.01
	ctree	0.15	0.13	0.10	0.13	0.13	0.16	0.10	0.08	-0.01

Continued on next page

Table D.11 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
AutoMPG	GOST	0.66	0.62	0.66	0.65	0.65	0.60	0.51	0.38	0.54
	rpart	0.73	0.62	0.69	0.69	0.57	0.63	0.58	0.51	0.40
	ctree	0.71	0.67	0.63	0.63	0.65	0.49	0.50	0.39	0.59
Behavior Urban Traffic	GOST	0.25	0.29	0.22	0.06	0.08	0.10	0.06	0.10	0.00
	rpart	0.25	0.29	0.22	0.06	0.23	0.14	0.06	0.10	0.09
	ctree	0.25	0.29	0.22	0.09	0.11	0.17	0.10	0.00	0.00
BikeSharing	GOST	0.97	0.93	0.94	0.94	0.93	0.94	0.93	0.93	0.94
	rpart	0.83	0.89	0.91	0.91	0.92	0.92	0.93	0.92	0.94
	ctree	0.95	0.94	0.94	0.93	0.95	0.95	0.94	0.95	0.95
Blog Feedback	GOST	-0.25	-0.24	-0.23	-0.24	-0.24	-0.24	1.00	1.00	1.00
	rpart	-0.25	-0.24	-0.23	-0.23	-0.23	-0.24	1.00	1.00	1.00
	ctree	-0.25	-0.25	-0.24	-0.23	-0.24	-0.24	1.00	1.00	1.00
Buzz in Social Media	GOST	0.81	0.80	0.78	0.77	0.73	0.63	0.28	1.00	1.00
	rpart	0.80	0.78	0.76	0.75	0.71	0.62	0.26	1.00	1.00
	ctree	0.81	0.80	0.78	0.76	0.72	0.60	0.24	1.00	1.00
Cargo2000	GOST	0.02	0.02	0.03	0.05	1.00	1.00	1.00	0.99	-2.10
	rpart	0.02	0.03	0.03	0.04	1.00	1.00	0.99	0.99	-2.07
	ctree	0.03	0.03	0.04	0.04	0.65	0.81	1.00	1.00	-2.10
Communities Crime	GOST	0.76	0.70	0.59	0.62	0.65	0.64	0.70	0.72	0.75
	rpart	0.58	0.70	0.69	0.69	0.71	0.76	0.77	0.71	0.69
	ctree	0.76	0.75	0.72	0.76	0.77	0.69	0.74	0.76	0.77
Computer Hardware	GOST	0.88	0.83	0.77	0.58	0.51	0.63	0.55	0.91	0.90
	rpart	0.87	0.42	0.76	0.69	0.72	0.50	0.51	0.91	0.73
	ctree	0.84	0.87	0.75	0.34	0.28	0.49	0.51	0.77	0.69
Concrete Slump	GOST	0.14	0.15	0.16	0.11	0.00	0.09	0.18	0.00	0.17
	rpart	0.14	0.15	0.16	0.11	0.11	0.09	0.02	0.27	0.11
	ctree	0.00	0.00	0.00	0.11	0.00	0.09	0.07	0.00	0.17
Concrete Strength	GOST	0.54	0.51	0.56	0.37	0.41	0.42	0.52	0.49	0.64
	rpart	0.39	0.43	0.54	0.39	0.47	0.44	0.48	0.54	0.58
	ctree	0.54	0.48	0.54	0.43	0.54	0.49	0.48	0.54	0.58
CSM	GOST	0.44	0.34	0.46	0.44	0.42	0.12	0.22	0.35	0.27
	rpart	0.45	0.49	0.46	0.48	0.45	0.43	0.35	0.35	0.32
	ctree	0.28	0.20	0.22	0.48	0.31	0.23	0.15	0.23	0.22
Cycle Power	GOST	0.84	0.81	0.78	0.75	0.73	0.74	0.73	0.71	0.63
	rpart	0.78	0.72	0.73	0.72	0.75	0.74	0.74	0.70	0.63
	ctree	0.84	0.82	0.76	0.74	0.74	0.75	0.73	0.70	0.63
Electrical Stability	GOST	0.50	0.50	0.46	0.49	0.44	0.42	0.42	0.38	0.32
	rpart	0.41	0.43	0.44	0.37	0.39	0.34	0.37	0.29	0.28
	ctree	0.46	0.48	0.44	0.47	0.45	0.45	0.44	0.40	0.34
Energy Efficiency 1	GOST	-0.91	-0.93	-0.91	-0.89	-0.65	0.96	0.81	0.75	0.79
	rpart	-0.51	-0.50	-0.92	-0.90	-0.88	0.95	0.88	0.75	0.74
	ctree	-0.91	-0.92	-0.91	-0.89	-0.88	0.94	0.85	0.75	0.75
Energy Efficiency 2	GOST	-1.06	-1.01	-1.06	-1.00	-0.99	0.99	1.00	0.97	0.92
	rpart	-0.44	-0.41	-1.06	-1.00	-1.00	1.00	0.95	0.97	0.88
	ctree	-1.06	-1.01	-1.05	-1.00	-1.00	0.97	0.94	0.91	0.86
Faceboook Comments	GOST	-0.44	-0.41	-0.41	-0.41	-0.40	-0.40	-0.40	1.00	1.00
	rpart	-0.42	-0.41	-0.42	-0.40	-0.40	-0.40	-0.39	1.00	1.00
	ctree	-0.43	-0.42	-0.43	-0.43	-0.42	-0.43	-0.43	1.00	1.00

Continued on next page

Table D.11 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Faceboook Metrics	GOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.19
	rpart	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.24	0.17
	ctree	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19
Fires	GOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	rpart	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	ctree	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
GeoMusic	GOST	0.02	0.00	0.02	0.00	-0.00	0.07	0.05	0.00	0.06
	rpart	0.11	0.02	0.06	0.09	0.07	0.07	0.05	0.00	0.06
	ctree	0.04	0.02	0.03	0.05	0.01	0.03	0.05	0.01	-0.01
Insurance Benchmark	GOST	-0.01	-0.00	-0.01	0.00	0.00	-0.00	1.00	1.00	1.00
	rpart	-0.01	-0.01	-0.01	-0.00	0.00	0.00	1.00	1.00	1.00
	ctree	-0.00	-0.00	-0.00	0.00	0.00	0.00	1.00	1.00	1.00
KEGG Directed	GOST	0.83	0.85	0.84	0.87	0.86	0.88	0.91	-2.77	-2.71
	rpart	0.80	0.83	0.85	0.85	0.86	0.85	0.87	-2.66	-2.65
	ctree	0.79	0.81	0.85	0.86	0.87	0.88	0.90	-2.68	-2.63
KEGG Undirected	GOST	0.89	0.86	0.87	0.87	0.82	0.90	0.87	0.89	0.87
	rpart	0.80	0.83	0.77	0.80	0.80	0.83	0.81	0.84	0.77
	ctree	0.84	0.83	0.85	0.85	0.82	0.91	0.88	0.87	0.81
Kernel Performance	GOST	0.62	0.60	0.58	0.56	0.55	0.52	0.48	0.45	0.37
	rpart	0.53	0.48	0.49	0.47	0.45	0.42	0.39	0.38	0.29
	ctree	0.56	0.55	0.53	0.53	0.52	0.46	0.40	0.40	0.30
Las Vegas Strip	GOST	0.00	0.00	0.00	-0.00	0.07	0.06	0.00	0.00	0.02
	rpart	-0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ctree	0.04	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	0.00
Online News Popularity	GOST	0.08	0.08	0.08	0.08	0.06	0.03	0.03	-0.00	1.00
	rpart	0.08	0.09	0.08	0.07	0.06	0.05	0.02	-0.00	1.00
	ctree	0.09	0.09	0.10	0.09	0.09	0.06	0.03	-0.00	1.00
Online Video Characteristics	GOST	0.77	0.77	0.78	0.76	0.75	0.73	0.75	0.77	0.80
	rpart	0.73	0.71	0.71	0.71	0.70	0.71	0.75	0.76	0.79
	ctree	0.69	0.75	0.76	0.75	0.74	0.76	0.78	0.79	0.78
Optical Network	GOST	0.73	0.56	0.60	0.61	0.44	0.34	1.00	1.00	1.00
	rpart	0.61	0.56	0.77	0.57	0.52	0.07	1.00	1.00	1.00
	ctree	0.74	0.72	0.67	0.50	0.37	0.24	1.00	1.00	1.00
Parkinson Telemonitoring	GOST	0.60	0.70	0.53	0.55	0.62	0.69	0.72	0.79	0.79
	rpart	0.52	0.43	0.49	0.52	0.54	0.62	0.52	0.76	0.55
	ctree	0.39	0.35	0.36	0.41	0.48	0.57	0.53	0.62	0.61
PM2.5 - Beijing	GOST	0.36	0.40	0.41	0.42	0.43	0.45	0.47	0.47	0.42
	rpart	0.35	0.37	0.38	0.40	0.41	0.36	0.42	0.45	0.42
	ctree	0.36	0.38	0.39	0.41	0.43	0.45	0.45	0.47	0.42
Propulsion Plant	GOST	0.71	0.75	0.68	0.69	0.68	0.68	0.69	0.71	0.71
	rpart	0.47	0.48	0.52	0.45	0.44	0.34	0.50	0.47	0.49
	ctree	0.42	0.50	0.50	0.36	0.27	0.27	0.25	0.32	0.00
Protein	GOST	0.32	0.33	0.29	0.32	0.33	0.34	0.30	0.32	0.36
	rpart	0.28	0.30	0.30	0.29	0.27	0.29	0.26	0.24	0.24
	ctree	0.30	0.29	0.28	0.30	0.28	0.26	0.25	0.26	0.24
Real Estate 1	GOST	0.54	0.63	0.59	0.62	0.69	0.60	0.62	0.55	0.48
	rpart	0.39	0.46	0.61	0.65	0.67	0.60	0.62	0.48	0.60
	ctree	0.53	0.49	0.67	0.65	0.62	0.61	0.57	0.50	0.42

Continued on next page

Table D.11 – continued from previous page

Dataset	Method	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Real Estate 2	GOST	0.84	0.84	0.81	0.79	0.82	0.79	0.83	0.82	0.92
	rpart	0.72	0.75	0.71	0.71	0.72	0.74	0.81	0.79	0.91
	ctree	0.75	0.76	0.76	0.76	0.76	0.77	0.78	0.75	0.89
ResidentialBuilding	GOST	0.80	0.69	0.63	0.74	0.63	0.71	0.52	0.52	0.44
	rpart	0.56	0.86	0.83	0.73	0.78	0.64	0.73	0.52	0.44
	ctree	0.80	0.87	0.79	0.80	0.70	0.74	0.66	0.46	0.34
Servo	GOST	0.62	0.61	0.59	0.23	0.57	0.74	0.47	0.49	0.05
	rpart	0.31	0.31	0.30	0.42	0.49	0.18	0.17	0.32	0.08
	ctree	0.60	0.31	0.30	0.23	0.21	0.18	0.17	0.13	0.00
Stockmarket Istanbul	GOST	0.26	0.29	0.20	0.26	0.29	0.13	0.08	0.08	0.01
	rpart	0.13	0.26	0.21	0.28	0.27	0.16	0.08	0.08	0.01
	ctree	0.25	0.25	0.30	0.27	0.23	0.18	0.15	0.18	0.02
Stock Portfolio	GOST	0.71	0.42	0.86	0.49	0.59	0.48	0.00	0.48	0.39
	rpart	0.60	0.00	0.86	0.01	0.59	0.48	0.16	0.48	0.39
	ctree	0.02	0.12	-0.16	0.37	0.59	0.44	0.65	0.66	0.41
Student Performance	GOST	0.08	0.12	0.11	0.09	0.00	0.00	0.08	0.09	0.04
	rpart	0.08	0.12	0.11	0.09	0.12	0.06	0.08	0.09	0.24
	ctree	0.08	0.12	0.11	0.10	0.12	0.11	0.08	0.09	0.07
WineQuality	GOST	0.10	0.10	0.22	0.24	-0.25	-0.26	-0.25	-0.27	0.01
	rpart	0.10	0.09	0.22	0.22	-0.28	-0.25	-0.25	-0.27	0.02
	ctree	0.11	0.11	0.24	0.25	-0.34	-0.35	-0.34	-0.36	0.01
Yacht	GOST	0.91	0.78	0.80	0.83	0.76	0.69	0.87	0.80	0.80
	rpart	0.90	0.62	0.62	0.82	0.76	0.80	0.73	0.80	0.80
	ctree	0.94	0.98	0.80	0.85	0.86	0.94	0.56	0.85	0.59

Bibliography

[1] O. Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978.

[2] Abbott. m-PIMA™ ANALYSER, 2022. URL: www.globalpointofcare.abbott/en/product-details/m-pima-analyser.html.

[3] A. Acharya, K. Cunningham, S. Manandhar, N. Shrestha, M. Chen, and A. Weissman. Exploring the use of mobile health to improve community-based health and nutrition service utilization in the hills of Nepal: qualitative study. *Journal of Medical Internet Research*, 22(9):1–9, 2020. doi:10.2196/17659.

[4] C. I. Aci and M. F. Akay. A hybrid congestion control algorithm for broadcast-based architectures with multiple input queues. *The Journal of Supercomputing*, 71(5):1907–1931, May 2015. doi:10.1007/s11227-015-1384-1.

[5] M. Ahmed, M. Jahangir, H. Afzal, A. Majeed, and I. Siddiqi. Using crowd-source based features from social media and conventional features to predict the movies popularity. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pages 273–278, Dec. 2015. doi:10.1109/SmartCity.2015.83.

[6] Airtel Malawi. Airtel integrated report 2019, 2020. URL: https://www.airtel.mw/assets/pdf/investor/Airtel-Malawi-PLC-Annual-Report-2019.pdf.

[7] A. Ak and A. L. Erera. A paired-vehicle recourse strategy for the vehicle-routing problem with stochastic demands. *Transportation Science*, 41(2):222–237, 2007.

[8] O. Akbilgic, H. Bozdogan, and M. E. Balaban. A novel hybrid RBF neural networks model as a forecaster. *Statistics and Computing*, 24, May 2013. doi:10.1007/s11222-013-9375-7.

[9] J. C. Aker and I. M. Mbiti. Mobile phones and economic development in rural Africa. *Journal of Economic Perspectives*, 24(3):207–232, 2010. doi:10.1080/00220388.2012.709615.

[10] M. Albareda-Sambola, E. Fernández, and G. Laporte. The dynamic multiperiod vehicle routing problem with probabilistic information. *Computers & Operations Research*, 48:31–39, 2014.

[11] G. A. Alemnji, C. Zeh, K. Yao, and P. N. Fonjungo. Strengthening national health laboratories in sub-Saharan Africa: a decade of remarkable progress. *Tropical Medicine & International Health*, 19(4):450–458, 2014.

[12] C. B. Aranda-Jan, N. Mohutsiwa-Dibe, and S. Loukanova. Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa. *BMC Public Health*, 14(1), 2014. doi:10.1186/1471-2458-14-188.

[13] C. Archetti, D. Feillet, and M. G. Speranza. Complexity of routing problems with release dates. *European Journal of Operational Research*, 247(3):797–803, 2015.

[14] J. B. Babigumira, S. J. Lubinga, M. M. Cheng, J. K. Karichu, and L. P. Garrison Jr. Potential cost-effectiveness of HIV viral load sample collection and testing methods in Malawi. *DOI:10.21203/rs.2.13295/v1*, 2019. doi:10.21203/rs.2.13295/v1.

[15] R. Baldacci, E. Bartolini, A. Mingozzi, and A. Valletta. An exact algorithm for the period routing problem. *Operations research*, 59(1):228–241, 2011.

[16] R. Ballester-Ripoll, E. G. Paredes, and R. Pajarola. Sobol tensor trains for global sensitivity analysis. *Reliability Engineering and System Safety*, 183, Dec. 2017. doi:10.1016/j.ress.2018.11.007.

[17] C. Banda, S. F. Nayupe, S. Munharo, P. Patel, and P. Mbulaje. Factors affecting viral load testing turnaround time in malawi. *PAMJ-One Health*, 6(5), 2021.

[18] Baobab Health Trust. Malawi Master Facility List. Accessed on 2022-04-02. URL: https://github.com/BaobabHealthTrust/master-facility-list/blob/master/health-facilities.csv.

[19] J. Barrington, O. Wereko-Brobby, P. Ward, W. Mwafongo, and S. Kungulwe. SMS for life: a pilot project to improve anti-malarial drug supply management in rural Tanzania using standard technology. *Malaria Journal*, 9(1):1–9, 2010. doi:10.1186/1475-2875-9-298.

[20] K. P. Bennett and J. Blue. Optimal decision trees. *Rensselaer Polytechnic Institute Math Report*, 214, 1996.

[21] G. Berbeglia, J.-F. Cordeau, and G. Laporte. Dynamic pickup and delivery problems. *European Journal of Operational Research*, 202(1):8–15, 2010.

[22] D. Bertsimas and J. Dunn. Optimal classification trees. *Machine Learning*, pages 1–44, 2017.

[23] D. Bertsimas, J. Dunn, E. Gibson, and A. Orfanoudaki. Optimal survival trees. *Machine Learning*, pages 1–73, 2022.

[24] D. Bertsimas and J. W. Dunn. *Machine Learning under a Modern Optimization Lens*. Dynamic Ideas LLC, 2019.

[25] D. J. Bertsimas. A vehicle routing problem with stochastic demand. *Operations Research*, 40(3):574–585, 1992.

[26] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: a fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. doi:10.1137/141000671.

[27] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine*, 17(10):1169–1186, 1998.

[28] A. Binagwaho, P. Mugwaneza, A. A. Irakoze, S. Nsanzimana, M. Agbonyitor, C. T. Nutt, C. M. Wagner, A. Rukundo, A. Ahayo, P. Drobac, C. Karema, R. Hinda, L. Leung, S. Bandara, E. Chopyak, and M. C. S. Fawzi. Scaling up early infant diagnosis of HIV in Rwanda, 2008–2010. *Journal of Public Health Policy*, 34(1):2–16, 2013. doi:10.1057/jphp.2012.62.

[29] C. E. Boeke, J. Joseph, C. Atem, C. Banda, K. D. Coulibaly, N. Doi, A. Gunda, J. Kandulu, B. Kiernan, L. Kingwara, et al. Evaluation of near point-of-care viral load implementation in public health facilities across seven countries in sub-Saharan Africa. *Journal of the International AIDS Society*, 24(1):e25663, 2021.

[30] N. Bostel, P. Dejax, P. Guez, and F. Tricoire. Multiperiod planning and routing on a rolling horizon for field force optimization logistics. In *The Vehicle Routing Problem: Latest Advances and New Challenges*, pages 503–525. Springer, 2008.

[31] I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011.

[32] J. J. Boutilier and T. C. Chan. Ambulance emergency response optimization in developing countries. 2018. *arXiv preprint arXiv:1801.05402*.

[33] J. J. Boutilier, J. O. Jónasson, and E. Yoeli. Improving TB treatment adherence support: the case for targeted behavioral interventions. 2020. *Working Paper*.

[34] L. Breiman. Software for the masses, 2002. URL: www.stat.berkeley.edu/~breiman/wald2002-3.pdf.

[35] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. CRC press, 1984.

[36] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, 78(1):1–3, 1950.

[37] M. A. Bulterys, P. Oyaro, E. Brown, N. Yongo, E. Karauki, J. Wagude, L. Kingwara, N. Bowen, S. Njogo, A. D. Wagner, et al. Costs of point-of-care viral load testing for adults and children living with HIV in Kenya. *Diagnostics*, 11(1):140, 2021.

[38] F. Cabitza, R. Rasoini, and G. F. Gensini. Unintended consequences of machine learning in medicine. *Journal of the American Medical Association*, 318(6):517–518, 2017.

[39] A. M. Campbell and B. W. Thomas. Challenges and advances in a priori routing. In *The Vehicle Routing Problem: Latest Advances and New Challenges*, pages 123–142. Springer, 2008.

[40] L. M. Candanedo, V. Feldheim, and D. Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140:81–97, 2017.

[41] D. Castelvecchi. Can we open the black box of AI? *Nature News*, 538(7623):20, 2016.

[42] D. Cattaruzza, N. Absi, and D. Feillet. The multi-trip vehicle routing problem with time windows and release dates. *Transportation Science*, 50(2):676–693, 2016.

[43] CDC. Malawi Country Profile, 2021. Accessed on 2022-06-21. URL: https://www.cdc.gov/globalhivtb/where-we-work/malawi/malawi.html.

[44] Cepheid, 2022. Accessed on 2022-07-15. URL: https://www.cepheid.com/en_US/systems/GeneXpert-Family-of-Systems/GeneXpert-System.

[45] P. A. Chou. Optimal partitioning for classification and regression trees. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (4):340–354, 1991.

[46] A. Ciampi, C.-H. Chang, S. Hogg, and S. McKinney. Recursive partition: a versatile method for exploratory data analysis in biostatistics. *Biostatistics*, pages 23–50, 1987.

[47] A. Ciampi, J. Thiffault, J. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*, 4(3):185–204, 1986.

[48] C. Cintron, V. Mudhune, R. Haider, et al. Costs of HIV viral load and early infant diagnosis testing in Kenya. *Health Finance and Governance Project, Abt Associates Inc*, 2017.

[49] A. Coraddu, L. Oneto, A. Ghio, S. Savio, D. Anguita, and M. Figari. Machine learning approaches for improving condition-based maintenance of naval propulsion plants. *Proceedings of the Institution of Mechanical Engineers, Part*

M: Journal of Engineering for the Maritime Environment, 230(1):136–153, 2016. arXiv:https://doi.org/10.1177/1475090214540874, doi:10.1177/1475090214540874.

[50] J.-F. Cordeau, G. Laporte, M. W. P. Savelsbergh, and D. Vigo. Vehicle routing. *Handbooks in Operations Research and Management Science*, 14:367–428, 2007.

[51] P. Cortez, A. Cerdeira, F. L. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47:547–553, 2009.

[52] B. Cox, M. Blaxter, A. Buckle, N. Fenner, J. Golding, M. Gore, F. Huppert, J. Nickson, S. M. Roth, J. Stark, et al. *The health and lifestyle survey. Preliminary report of a nationwide survey of the physical and mental health, attitudes and lifestyle of a random sample of 9,003 British adults.* Health Promotion Research Trust, 1987.

[53] D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society*, 34:187–220, 1972.

[54] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[55] R. B. D'Agostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro, and W. B. Kannel. General cardiovascular risk profile for use in primary care. *Circulation*, 117, 2008.

[56] R. B. Davis and J. R. Anderson. Exponential survival trees. *Statistics in Medicine*, 8(8):947–961, 1989.

[57] K. De Boeck, C. Decouttere, and N. Vandaele. Vaccine distribution chains in low- and middle-income countries: a literature review. *Omega*, page 102097, 2019.

[58] C. Déglise, L. Suzanne Suggs, and P. Odermatt. SMS for disease control in developing countries: a systematic review of mobile health applications. *Journal of Telemedicine and Telecare*, 18(5):273–281, 2012. doi:10.1258/jtt.2012.110810.

[59] S. Deo, L. Crea, J. Quevedo, J. Lehe, L. Vojnov, T. Peter, and I. Jani. Expedited results delivery systems using GPRS technology significantly reduce early infant diagnosis test turnaround times. *Journal of Acquired Immune Deficiency Syndromes*, 70(1):e1–e4, 2015.

[60] S. Deo and M. Sohoni. Optimal decentralization of early infant diagnosis of HIV in resource-limited settings. *Manufacturing & Service Operations Management*, 17(2):191–207, 2015. doi:10.1287/msom.2014.0512.

[61] Diagnostics for the Real World, 2022. URL: drw-ltd.com/products.html.

[62] P. K. Drain, J. Dorward, L. R. Violette, J. Quame-Amaglo, K. K. Thomas, N. Samsunder, H. Ngobese, K. Mlisana, P. Moodley, D. Donnell, et al. Point-of-care HIV viral load testing combined with task shifting to improve treatment outcomes (stream): findings from an open-label, non-inferiority, randomised controlled trial. *The Lancet HIV*, 2020.

[63] D. Dua and C. Graff. UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.

[64] J. Dunn. *Optimal Trees for Prediction and Prescription.* PhD thesis, Massachusetts Institute of Technology, 2018.

[65] I. Dunning, J. Huchette, and M. Lubin. JuMP: a modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017.

[66] L. Evers and C.-M. Messow. Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, 24(14):1632–1638, 2008.

[67] Expert Panel on Detection and Evaluation and Treatment of High Blood Cholesterol in Adults. Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). *Journal of the American Medical Association*, 285, 2001.

[68] H. Fanaee-T and J. Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3):113–127, 2014.

[69] T. Faruna, E. Akintunde, and B. Odelola. Leveraging private sector transportation/logistics services to improve the National Integrated Specimen Referral Network in Nigeria. *Business Management Dynamics*, 8(7):8, 2019.

[70] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019.

[71] FIND, 2022. Accessed on 2022-06-22. URL: https://www.finddx.org/pricing/genexpert/.

[72] S. Fotso. Deep neural networks for survival analysis based on a multi-task framework. 2018. arXiv:1801.05512.

[73] C. J. K. Fouodo, I. R. König, C. Weihs, A. Ziegler, and M. N. Wright. Support vector machines for survival analysis with R. *R Journal*, 10(1), 2018.

[74] S. C. Frank, J. Cohn, L. Dunning, E. Sacks, R. P. Walensky, S. Mukherjee, C. M. Dugdale, E. Turunga, K. A. Freedberg, and A. L. Ciaranello. Clinical effect and cost-effectiveness of incorporation of point-of-care assays into early infant HIV diagnosis programmes in Zimbabwe: a modelling study. *The Lancet HIV*, 6(3):e182–e190, 2019.

[75] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics New York, 2001.

[76] J. Gallien, I. Rashkova, R. Atun, and P. Yadav. National drug stockout risks and the global fund disbursement process for procurement. *Production and Operations Management*, 26(6):997–1014, 2017.

[77] P. Ganesh, T. Heller, B. Chione, J. Gumulira, S. Gugsa, S. Khan, S. McGovern, A. Nhlema, L. Nkhoma, J. A. Sacks, et al. Near point-of-care HIV viral load: targeted testing at large facilities. *Journal of Acquired Immune Deficiency Syndromes*, 86(2):258, 2021.

[78] E. Gibson, S. Deo, J. O. Jónasson, M. Kachule, and K. Palamountain. Redesigning sample transportation in Malawi through improved data sharing and daily route optimization. *SSRN Electronic Journal*, 2020. doi:10.2139/ssrn.3712556.

[79] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: an overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. IEEE, 2018.

[80] S. J. Girdwood, T. Crompton, M. Sharma, J. Dorward, N. Garrett, P. K. Drain, W. Stevens, and B. E. Nichols. Cost-effectiveness of adoption strategies for point of care HIV viral load monitoring in South Africa. *EClinicalMedicine*, 28:100607, 2020. doi:10.1016/j.eclinm.2020.100607.

[81] S. J. Girdwood, B. E. Nichols, C. Moyo, T. Crompton, D. Chimhamhiwa, and S. Rosen. Optimizing viral load testing access for the last mile: geospatial cost model for point of care instrument placement. *PLOS ONE*, 14(8):e0221586, 2019.

[82] E. Giunchiglia, A. Nemchenko, and M. van der Schaar. RNN-SURV: a deep recurrent model for survival analysis. In *International Conference on Artificial Neural Networks*, pages 23–32. Springer, 2018.

[83] B. L. Golden, S. Raghavan, and E. A. Wasil. *The Vehicle Routing Problem: Latest Advances and New Challenges*, volume 43. Springer Science & Business Media, 2008.

[84] J. C. Goodson, J. W. Ohlmann, and B. W. Thomas. Rollout policies for dynamic solutions to the multivehicle routing problem with stochastic demand and duration limits. *Operations Research*, 61(1):138–154, 2013.

[85] L. Gordon and R. A. Olshen. Tree-structured survival analysis. *Cancer treatment reports*, 69(10):1065–1069, 1985.

[86] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

[87] T. Grubinger, A. Zeileis, and K.-P. Pfeiffer. evtree: evolutionary learning of globally optimal classification and regression trees in R. *Journal of Statistical Software*, 61(1):1–29, 2014. URL: https://www.jstatsoft.org/v061/i01, doi:10.18637/jss.v061.i01.

[88] GSMA Association. The Mobile Economy: Sub-Saharan Africa, 2019. URL: https://www.gsmaintelligence.com/research/2019/02/the-mobile-economy-2019/731/.

[89] Gurobi Optimization, LLC. Gurobi optimizer reference manual, 2022. URL: https://www.gurobi.com.

[90] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546, 1982.

[91] J. Hattie and H. Timperley. The power of feedback. *Review of Educational Research*, 77(1):81–112, 2007. doi:10.3102/003465430298487.

[92] P. Herd, D. Carr, and C. Roan. Cohort profile: Wisconsin longitudinal study (WLS). *International Journal of Epidemiology*, 43(1):34–41, 2014.

[93] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2005.

[94] T. Hothorn, K. Hornik, C. Strobl, and A. Zeileis. Party: a laboratory for recursive partitioning, 2010. URL: https://cran.r-project.org/web/packages/party/vignettes/party.pdf.

[95] T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

[96] T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Tröger. Bagging survival trees. *Statistics in Medicine*, 23(1):77–91, 2004.

[97] IHME. Global Burden of Disease Study 2019 Data Resources, 2019. Accessed on 2022-06-21. URL: https://ghdx.healthdata.org/gbd-2019.

[98] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

[99] J. O. Jónasson, S. Deo, and J. Gallien. Improving HIV early infant diagnosis supply chains in sub-Saharan Africa: models and application to Mozambique. *Operations Research*, 65(6):1479–1493, 2017. doi:10.1287/opre.2017.1646.

[100] S. Kamaljot, S. Ranjeet Kaur, and D. Kumar. Comment volume prediction using neural networks and decision trees. In *IEEE 17th UKSIM-AMSS International Conference on Modelling and Simulation*, Cambridge, United Kingdom, Mar. 2015.

[101] S. Kanters, M. Vitoria, M. Zoratti, M. Doherty, M. Penazzato, A. Rangaraj, N. Ford, K. Thorlund, A. H. Anis, M. E. Karim, et al. Comparative efficacy, tolerability and safety of dolutegravir and efavirenz 400mg among antiretroviral therapies for first-line hiv treatment: a systematic literature review and network meta-analysis. *EClinicalMedicine*, 28:100573, 2020.

[102] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[103] M. Kaul, B. Yang, and C. S. Jensen. Building accurate 3d spatial networks to enable next generation intelligent transportation systems. In *IEEE 14th International Conference on Mobile Data Management*, volume 1, pages 137–146. IEEE, 2013.

[104] F. Kawala, A. Douzal-Chouakria, E. Gaussier, and E. Dimert. Prédictions d'activité dans les réseaux sociaux en ligne. In *4ième conférence sur les modèles et l'analyse des réseaux : Approches mathématiques et informatiques*, page 16, France, Oct. 2013. URL: https://hal.archives-ouvertes.fr/hal-00881395.

[105] K. Kayumba, S. Nsanzimana, A. Binagwaho, P. Mugwaneza, J. Rusine, E. Remera, J. B. Koama, V. Ndahindwa, P. Johnson, D. J. Riedel, and J. Condo. TRACnet internet and short message service technology improves time to antiretroviral therapy initiation among HIV-infected infants in Rwanda. *The Pediatric Infectious Disease Journal*, 35(7):767–771, July 2016. URL: http://journals.lww.com/00006454-201607000-00011, doi:10.1097/INF.0000000000001153.

[106] Y. Kebede, P. N. Fonjungo, G. Tibesso, R. Shrivastava, J. N. Nkengasong, T. Kenyon, A. Kebede, R. Gadde, and G. Ayana. Improved specimen-referral system and increased access to quality laboratory services in Ethiopia: the role of the public-private partnership. *The Journal of Infectious Diseases*, 213(Supplement_2):S59–S64, 2016.

[107] C. Keeton. Measuring the impact of e-health. *Bulletin of the World Health Organization*, 90(5):326–327, 2012. doi:10.2471/blt.12.020512.

[108] D. Killian, E. Gibson, M. Kachule, K. Palamountain, J. B. Bangoh, S. Deo, J. O. Jonasson, et al. An unstructured supplementary service data system for daily tracking of patient samples and diagnostic results in a diagnostic network in malawi: system development and field trial. *Journal of Medical Internet Research*, 23(7):e26582, 2021.

[109] C. Kiyaga, H. Sendagire, E. Joseph, I. McConnell, J. Grosz, V. Narayan, G. Esiru, P. Elyanu, Z. Akol, W. Kirungi, J. Musinguzi, and A. Opio. Uganda's new national laboratory sample transport system: a successful model for improving access to diagnostic services for early infant HIV diagnosis and other programs. *PLOS ONE*, 8(11):1–7, 2013. doi:10.1371/journal.pone.0078609.

[110] M. A. Klapp, A. L. Erera, and A. Toriello. The one-dimensional dynamic dispatch waves problem. *Transportation Science*, 52(2):402–415, 2018.

[111] S. Kraiselburd and P. Yadav. Supply chains and global health: an imperative for bringing operations management scholarship into action. *Production and Operations Management*, 22(2):377–381, 2013.

[112] A. S. Laar, E. Bekyieriya, S. Isang, and B. Baguune. Assessment of mobile health technology for maternal and child health services in rural Upper West Region of Ghana. *Public Health*, 168(March):1–8, 2019. doi:10.1016/j.puhe.2018.11.014.

[113] A. J. Lankowski, M. J. Siedner, D. R. Bangsberg, and A. C. Tsai. Impact of geographic and transportation-related barriers on HIV outcomes in sub-Saharan Africa: a systematic review. *AIDS and Behavior*, 18(7):1199–1223, 2014. doi:10.1007/s10461-014-0729-8.

[114] H. Laurent and R. L. Rivest. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17, 1976.

[115] M. LeBlanc and J. Crowley. Relative risk trees for censored survival data. *Biometrics*, pages 411–425, 1992.

[116] M. LeBlanc and J. Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467, 1993.

[117] S. Lecher, D. Ellenberger, A. A. Kim, P. N. Fonjungo, S. Agolory, M. Y. Borget, L. Broyles, S. Carmona, G. Chipungu, K. M. De Cock, V. Deyde, M. Downer, S. Gupta, J. E. Kaplan, C. Kiyaga, N. Knight, W. MacLeod, B. Makumbi, H. Muttai, C. Mwangi, J. W. Mwangi, M. Mwasekaga, L. W. Ng'Ang'A, Y. Pillay, A. Sarr, S. Sawadogo, D. Singer, W. Stevens, C. A. Toure, and J. Nkengasong. Scale-up of HIV viral load monitoring — seven sub-Saharan African countries. *Morbidity and Mortality Weekly Report*, 64(46):1287–1290, 2015. doi:10.15585/mmwr.mm6446a3.

[118] S. L. Lecher, P. Fonjungo, D. Ellenberger, C. A. Toure, G. Alemnji, N. Bowen, F. Basiye, A. Beukes, S. Carmona, M. de Klerk, et al. HIV viral load monitoring among patients receiving antiretroviral therapy – eight sub-Saharan African countries, 2013–2018. *Morbidity and Mortality Weekly Report*, 70(21):775, 2021.

[119] N. V. Lemay, T. Sullivan, B. Jumbe, and C. P. Perry. Reaching remote health workers in Malawi: baseline assessment of a pilot mHealth intervention. *Journal of Health Communication*, 17(SUPPL. 1):105–117, 2012. doi:10.1080/10810730.2011.649106.

[120] N.-H. Z. Leung, A. Chen, P. Yadav, and J. Gallien. The impact of inventory management on stock-outs of essential drugs in sub-Saharan Africa: secondary analysis of a field experiment in Zambia. *PLOS ONE*, 11(5), 2016.

[121] R. Levi, G. Perakis, and G. Romero. On the effectiveness of uniform subsidies in increasing market consumption. *Management Science*, 63(1):40–57, 2017.

[122] X. Liang, S. Li, S. Zhang, H. Huang, and S. X. Chen. PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. *Journal of Geophysical Research: Atmospheres*, 121(17):10,220–10,236, 2016. doi:10.1002/2016JD024877.

[123] K. Liestbl, P. K. Andersen, and U. Andersen. Survival analysis and neural nets. *Statistics in Medicine*, 13(12):1189–1200, 1994.

[124] Y.-C. Liu and I.-C. Yeh. Using mixture design and neural networks to build stock selection decision support systems. *Neural Computing and Applications*, 28, Nov. 2015. doi:10.1007/s00521-015-2090-x.

[125] L. C. Long, M. Maskew, A. T. Brennan, C. Mongwenyana, C. Nyoni, G. Malete, I. Sanne, M. P. Fox, and S. Rosen. Initiating antiretroviral therapy for HIV at a patient's first clinic visit: a cost-effectiveness analysis of the rapid initiation of treatment randomized controlled trial. *AIDS*, 31(11):1611–1619, 2017.

[126] Malawi Ministry of Health. Malawi EID and Viral Load dashboard. Accessed on 2022-07-02. URL: http://www.eidmalawi.org/.

[127] Malawi Ministry of Health. Malawi Master Facility Registry. Accessed on 2022-04-02. URL: http://zipatala.health.gov.mw/facilities.

[128] Malawi Ministry of Health. Malawi guidelines for clinical management of HIV in children and adults, 4th edition, 2018. URL: https://differentiatedservicedelivery.org/Portals/0/adam/Content/yb4xSSLvEOSW98_z7wTm_w/File/MalawiClinicalHIVGuidelines2018(1).pdf.

[129] Malawi Ministry of Health. Addendum to the 4$^{\text{th}}$ edition of the Malawi integrated guidelines and standard operating procedures for clinical HIV services, 2019.

[130] Malawi National Statistical Office and ICF. Malawi demographic and health survey 2015–16, 2017.

[131] Malawi National Statistical Office and National Statistical Office of Malawi, United Nations Population Fund. Malawi Population and Housing Census, 2018. URL: http://www.nsomalawi.mw/index.php?option=com_content&view=article&id=226&Itemid=6.

[132] I. Medhi, S. Patnaik, E. Brunskill, S. N. N. Gautama, W. Thies, and K. Toyama. Designing mobile interfaces for novice and low-literacy users. *ACM Transactions on Computer-Human Interaction*, 18(1):1–28, 2011. doi:10.1145/1959022.1959024.

[133] A. Metzger, P. Leitner, D. Ivanović, E. Schmieders, R. Franklin, M. Carro, S. Dustdar, and K. Pohl. Comparing and combining predictive business process monitoring techniques. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(2):276–290, Feb. 2015. doi:10.1109/TSMC.2014.2347265.

[134] C. E. Miller, A. W. Tucker, and R. A. Zemlin. Integer programming formulation of traveling salesman problems. *Journal of the ACM*, 7(4):326–329, 1960.

[135] P. A. Minchella, G. Chipungu, A. A. Kim, A. Sarr, H. Ali, R. Mwenda, J. N. Nkengasong, and D. Singer. Specimen origin, type and testing laboratory are linked to longer turnaround times for HIV viral load testing in Malawi. *PLOS ONE*, 12(2):e0173009, 2017.

[136] A. Mohamed, A. Rizaner, and A. H. Ulusoy. Using data mining to predict instructor performance. *Procedia Computer Science*, 102:137–142, 2016. 12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS 2016, 29-30 August 2016, Vienna, Austria. URL: http://www.sciencedirect.com/science/article/pii/S1877050916325601, doi:10.1016/j.procs.2016.09.380.

[137] A. M. Molinaro, S. Dudoit, and M. J. Van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1):154–177, 2004.

[138] S. Moro, P. Rita, and J. Coelho. Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. *Tourism Management Perspectives*, 23:41 – 52, 2017. URL: http://www.sciencedirect.com/science/article/pii/S2211973617300387, doi:https://doi.org/10.1016/j.tmp.2017.04.003.

[139] S. Moro, P. Rita, and B. Vala. Predicting social media performance metrics and evaluation of the impact on brand building: a data mining approach. *Journal of Business Research*, 69(9):3341–3351, 2016. URL: http://www.sciencedirect.com/science/article/pii/S0148296316000813, doi:10.1016/j.jbusres.2016.02.010.

[140] MSF Access Campaign. Putting HIV and HCV to the test, 2017.

[141] L. Msimango, A. Gibbs, H. Shozi, H. Ngobese, H. Humphries, P. K. Drain, N. Garrett, and J. Dorward. Acceptability of point-of-care viral load testing to facilitate differentiated care: a qualitative assessment of people living with hiv and nurses in south africa. *BMC Health Services Research*, 20(1):1–9, 2020.

[142] S. Mukherjee, J. Cohn, A. L. Ciaranello, E. Sacks, O. Adetunji, A. Chadambuka, H. Mafaune, M. Makayi, N. McCann, and E. Turunga. Estimating the cost of point-of-care early infant diagnosis in a program setting: a case study using Abbott m-PIMA and Cepheid GeneXpert IV in Zimbabwe. *Journal of Acquired Immune Deficiency Syndromes*, 84:S63–S69, 2020.

[143] N. Narodytska, A. Ignatiev, F. Pereira, and J. Marques-Silva. Learning optimal decision trees with SAT. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 1362–1368, 2018.

[144] K. V. Natarajan and J. M. Swaminathan. Inventory management in humanitarian operations: impact of amount, schedule, and uncertainty in funding. *Manufacturing & Service Operations Management*, 16(4):595–603, 2014.

[145] P. Ndarukwa, M. J. Chimbari, and E. Sibanda. Assessment of levels of asthma control among adult patients with asthma at Chitungwiza Central Hospital, Zimbabwe. *Allergy, Asthma and Clinical Immunology*, 16(1):1–7, 2020. doi:10.1186/s13223-020-0405-7.

[146] W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

[147] M.-L. Newell, H. Coovadia, M. Cortina-Borja, N. Rollins, P. Gaillard, and F. Dabis. Mortality of infected and uninfected infants born to HIV-infected mothers in Africa: a pooled analysis. *The Lancet*, 364:1236–1243, 2004.

[148] J. A. Nhavoto and Å. Grönlund. Mobile technologies and geographic information systems to improve health care systems: a literature review. *JMIR mHealth and uHealth*, 2(2):e21, 2014. doi:10.2196/mhealth.3216.

[149] S. Nicholas, E. Poulet, L. Wolters, J. Wapling, A. Rakesh, I. Amoros, E. Szumilin, M. Gueguen, and B. Schramm. Point-of-care viral load monitoring: outcomes from a decentralized HIV programme in Malawi. *Journal of the International AIDS Society*, 22(8):1–9, 2019. doi:10.1002/jia2.25387.

[150] B. E. Nichols, S. J. Girdwood, T. Crompton, L. Stewart-Isherwood, L. Berrie, D. Chimhamhiwa, C. Moyo, J. Kuehnle, W. Stevens, S. Rosen, et al. Monitoring viral load for the last mile: what will it cost? *Journal of the International AIDS Society*, 22(9):e25337, 2019.

[151] B. E. Nichols, S. J. Girdwood, A. Shibemba, S. Sikota, C. J. Gill, L. Mwananyanda, L. Noble, L. Stewart-Isherwood, L. Scott, S. Carmona, et al. Cost and impact of dried blood spot versus plasma separation card for scale-up

of viral load testing in resource-limited settings. *Clinical Infectious Diseases*, 70(6):1014–1020, 2020.

[152] S. Nijssen and E. Fromont. Mining optimal decision trees from itemset lattices. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 530–539, 2007.

[153] S. Nijssen and E. Fromont. Optimal constraint-based decision tree induction from itemset lattices. *Data Mining and Knowledge Discovery*, 21(1):9–51, 2010.

[154] A. C. Noordam, A. George, A. B. Sharkey, A. Jafarli, S. S. Bakshi, and J. C. Kim. Assessing scale-up of mHealth innovations based on intervention complexity: two case studies of child health programs in Malawi and Zambia. *Journal of Health Communication*, 20(3):343–353, 2015. doi:10.1080/10810730.2014.965363.

[155] J. J. Nutor, H. O. Duah, P. Agbadi, P. A. Duodu, and K. W. Gondwe. Spatial analysis of factors associated with hiv infection in malawi: indicators for effective prevention. *BMC Public Health*, 20(1):1–14, 2020.

[156] W. O. Ochieng, T. Ye, C. Scheel, A. Lor, J. Saindon, S. L. Yee, M. I. Meltzer, V. Kapil, and K. Karem. Uncrewed aircraft systems versus motorcycles to deliver laboratory samples in West Africa: a comparative economic study. *The Lancet Global Health*, 8(1):e143–e151, 2020.

[157] M. Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.

[158] H. Parvin, S. Beygi, J. E. Helm, P. S. Larson, and M. P. Van Oyen. Distribution of medication considering information, transshipment, and clustering: malaria in Malawi. *Production and Operations Management*, 27(4):774–797, 2018.

[159] S. Patnaik, E. Brunskill, and W. Thies. Evaluating the accuracy of data collection on mobile phones: a study of forms, SMS, and voice. In *2009 International Conference on Information and Communication Technologies and Development*, pages 74–84, 2009. doi:10.1109/ICTD.2009.5426700.

[160] PEPFAR. PEPFAR Panorama Spotlight. Accessed 2022-06-24. URL: https://data.pepfar.gov/library.

[161] PEPFAR. Malawi country operational plan 2019, 2019.

[162] PEPFAR. Malawi country operational plan 2021 strategic direction summary, 2021.

[163] T. Perrier, B. DeRenzi, and R. Anderson. USSD: the third universal app. In *Proceedings of the 2015 Annual Symposium on Computing for Development*, pages 13–21, New York, New York, USA, 2015. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2830629.2830645, doi:10.1145/2830629.2830645.

[164] T. Peter, C. Zeh, Z. Katz, A. Elbireer, B. Alemayehu, L. Vojnov, A. Costa, N. Doi, and I. Jani. Scaling up HIV viral load — lessons from the large-scale implementation of HIV early infant diagnosis and CD4 testing. *Journal of the International AIDS Society*, 20(S7):e25008, 2017.

[165] D. H. Peters, T. Adam, O. Alonge, I. A. Agyepong, and N. Tran. Republished research: implementation research: what it is and how to do it. *British Journal of Sports Medicine*, 48(8):731–736, 2014. doi:10.1136/bmj.f6753.

[166] M. D. Pham, L. Romero, B. Parnell, D. A. Anderson, S. M. Crowe, and S. Luchters. Feasibility of antiretroviral treatment monitoring in the era of decentralized HIV care: a systematic review. *AIDS Research and Therapy*, 14(3), 2017.

[167] PHIA News. Second population survey assessing HIV in Malawi shows progress toward epidemic control. URL: https://phia.icap.columbia.edu/second-population-survey-assessing-hiv-in-malawi-shows-progress-toward-epidemic-control.

[168] A. Phillips. HIV Synthesis Model. Accessed on 2022-07-05. URL: https://github-pages.ucl.ac.uk/hiv-synthesis/.

[169] A. Phillips, A. Shroufi, L. Vojnov, J. Cohn, T. Roberts, T. Ellman, K. Bonner, C. Rousseau, G. Garnett, V. Cambiano, et al. Sustainable HIV treatment in Africa through viral-load-informed differentiated care. *Nature*, 528(7580):S68–S76, 2015.

[170] A. N. Phillips, L. Bansi-Matharu, V. Cambiano, P. Ehrenkranz, C. Serenata, F. Venter, S. Pett, C. Flexner, A. Jahn, P. Revill, et al. The potential role of long-acting injectable cabotegravir–rilpivirine in the treatment of HIV in sub-Saharan Africa: a modelling analysis. *The Lancet Global Health*, 9(5):e620–e627, 2021.

[171] A. N. Phillips, L. Bansi-Matharu, F. Venter, D. Havlir, A. Pozniak, D. R. Kuritzkes, A. Wensing, J. D. Lundgren, D. Pillay, J. Mellors, et al. Updated assessment of risks and benefits of dolutegravir versus efavirenz in new antiretroviral treatment initiators in sub-Saharan Africa: modelling to inform treatment guidelines. *The Lancet HIV*, 7(3):e193–e200, 2020.

[172] A. N. Phillips, V. Cambiano, F. Nakagawa, L. Bansi-Matharu, D. Wilson, I. Jani, T. Apollo, M. Sculpher, T. Hallett, C. Kerr, et al. Cost-per-diagnosis as a metric for monitoring cost-effectiveness of HIV testing programmes in low-income settings in southern Africa: health economic and modelling analysis. *Journal of the International AIDS Society*, 22(7):e25325, 2019.

[173] A. N. Phillips, V. Cambiano, F. Nakagawa, D. Ford, T. Apollo, J. Murungu, C. Rousseau, G. Garnett, P. Ehrenkranz, L. Bansi-Matharu, et al. Point-of-care

viral load testing for sub-Saharan Africa: informing a target product profile. In *Open Forum Infectious Diseases*, volume 3. Oxford University Press, 2016.

[174] A. N. Phillips, V. Cambiano, F. Nakagawa, P. Revill, M. R. Jordan, T. B. Hallett, M. Doherty, A. De Luca, J. D. Lundgren, M. Mhangara, et al. Cost-effectiveness of public-health policy options in the presence of pretreatment NNRTI drug resistance in sub-Saharan Africa: a modelling study. *The Lancet HIV*, 5(3):e146–e154, 2018.

[175] A. N. Phillips, F. Venter, D. Havlir, A. Pozniak, D. Kuritzkes, A. Wensing, J. D. Lundgren, A. De Luca, D. Pillay, J. Mellors, et al. Risks and benefits of dolutegravir-based antiretroviral drug regimens in sub-Saharan Africa: a modelling study. *The Lancet HIV*, 6(2):e116–e127, 2019.

[176] V. Pillac, M. Gendreau, C. Guéret, and A. L. Medaglia. A review of dynamic vehicle routing problems. *European Journal of Operational Research*, 225(1):1–11, 2013.

[177] E. Proctor, H. Silmere, R. Raghavan, P. Hovmand, G. Aarons, A. Bunger, R. Griffey, and M. Hensley. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(2):65–76, 2011. doi:10.1007/s10488-010-0319-7.

[178] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[179] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Elsevier, 2014.

[180] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL: https://www.R-project.org/.

[181] M. Radespiel-Tröger, T. Rabenstein, H. T. Schneider, and B. Lausen. Comparison of tree-based methods for prognostic stratification of survival data. *Artificial Intelligence in Medicine*, 28(3):323–341, 2003.

[182] M. H. Rafiei and H. Adeli. A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management*, 142(2):04015066, 2016. URL: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29CO.1943-7862.0001047, arXiv:https://ascelibrary.org/doi/pdf/10.1061/\%28ASCE\%29CO.1943-7862.0001047, doi:10.1061/(ASCE)CO.1943-7862.0001047.

[183] A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[184] A. Reddy and L.-P. Kronek. Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data. *Bioinformatics*, 24(16):i248–i253, 08 2008. doi:10.1093/bioinformatics/btn265.

[185] M. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

[186] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz. Internal versus external cluster validation indexes. *International Journal of computers and communications*, 5(1):27–34, 2011.

[187] B. D. Ripley and R. M. Ripley. Neural networks as statistical methods in survival analysis. *Clinical applications of artificial neural networks*, pages 237–255, 2001.

[188] T. Roberts, J. Cohn, K. Bonner, and S. Hargreaves. Scale-up of routine viral load testing in resource-poor settings: current and future implementation challenges. *Clinical Infectious Diseases*, 62(8):1043–1048, 2016. arXiv:http://oup.prod.sis.lan/cid/article-pdf/62/8/1043/7450770/ciw001.pdf.

[189] S. E. Rutstein, M. C. Hosseinipour, D. Kamwendo, A. Soko, M. Mkandawire, A. K. Biddle, W. C. Miller, M. Weinberger, S. B. Wheeler, A. Sarr, et al. Dried blood spots for viral load monitoring in Malawi: feasible and effective. *PLOS ONE*, 10(4):e0124748, 2015.

[190] W. Samek and K.-R. Müller. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 5–22. Springer, 2019.

[191] A. Schöpperle. *Analysis of Challenges of Medical Supply Chains in Sub-Saharan Africa Regarding Inventory Management and Transport and Distribution*. Project thesis, University of Westminster, 2013. URL: http://www.transaid.org/medical-supply-chain-challenges---masterthesis.

[192] C. Scott and R. D. Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 52(4):1335–1353, 2006.

[193] N. Secomandi and F. Margot. Reoptimization approaches for the vehicle-routing problem with stochastic demands. *Operations Research*, 57(1):214–230, 2009.

[194] M. Shieshia, M. Noel, S. Andersson, B. Felling, S. Alva, S. Agarwal, A. Lefevre, A. Misomali, B. Chimphanga, H. Nsona, and Y. Chandani. Strengthening community health supply chain performance through an integrated approach: using mHealth technology and multilevel teams in Malawi. *Journal of Global Health*, 4(2), 2014. doi:10.7189/jogh.04.020406.

[195] M. J. Siedner, D. Santorino, A. J. Lankowski, M. Kanyesigye, M. B. Bwana, J. E. Haberer, and D. R. Bangsberg. A combination SMS and transportation reimbursement intervention to improve HIV care following abnormal CD4 test results in rural Uganda: a prospective observational cohort study. *BMC Medicine*, 13(1), 2015. doi:10.1186/s12916-015-0397-1.

[196] K. Simeon, M. Sharma, J. Dorward, J. Naidoo, N. Dlamini, P. Moodley, N. Samsunder, R. V. Barnabas, N. Garrett, and P. K. Drain. Comparative cost analysis of point-of-care versus laboratory-based testing to initiate and monitor HIV treatment in South Africa. *PLOS ONE*, 14(10):1–16, 2019. doi:10.1371/journal.pone.0223669.

[197] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1, 2011.

[198] A. Smith, A. Colmant, L. Oosthuizen, and J. H. van Vuuren. A new vehicle routing problem with application to pathology laboratory service delivery. In *44th Annual Conference of the Operations Research Society of South Africa*, page 72, 2015.

[199] N. H. Son. From optimal hyperplanes to optimal decision trees. *Fundamenta Informaticae*, 34(1, 2):145–174, 1998.

[200] W. S. Stevens and T. M. Marshall. Challenges in implemeting HIV load testing in South Africa. *Journal of Infectious Diseases*, 201(Supplement_1):S78–S84, 2010.

[201] A. Subramanyam, F. Mufalli, J. M. Laínez-Aguirre, J. M. Pinto, and C. E. Gounaris. Robust multiperiod vehicle routing under customer order uncertainty. *Operations Research*, 69(1):30–60, 2021.

[202] S.-C. Suen, D. Negoescu, and J. Goh. Design of incentive programs for optimal medication adherence. 2020. *Available at SSRN 3308510.*

[203] C. G. Sutcliffe, N. Moyo, J. L. Schue, J. N. Mutanga, M. Hamahuwa, P. Munachoonga, S. Maunga, P. E. Thuma, and W. J. Moss. The NSEBA demonstration project: implementation of a point-of-care platform for early infant diagnosis of HIV in rural Zambia. *Tropical Medicine & International Health*, 26(9):1036–1046, 2021.

[204] C. G. Sutcliffe, P. E. Thuma, J. H. van Dijk, K. Sinywimaanzi, S. Mweetwa, M. Hamahuwa, and W. J. Moss. Use of mobile phones and text messaging to decrease the turnaround time for early infant HIV diagnosis and notification in rural Zambia: an observational study. *BMC Pediatrics*, 17(1):1–9, 2017. doi:10.1186/s12887-017-0822-z.

[205] T. A. Taylor and W. Xiao. Subsidizing the distribution channel: donor funding to improve the availability of malaria drugs. *Management Science*, 60(10):2461–2477, 2014.

[206] T. M. Therneau, B. Atkinson, and B. Ripley. The rpart package, 2010.

[207] T. M. Therneau, P. M. Grambsch, and T. R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[208] R. Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

[209] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4):884–893, Apr. 2010. doi:10.1109/TBME.2009.2036000.

[210] A. Tsanas and A. Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560 – 567, 2012. URL: http://www.sciencedirect.com/science/article/pii/S037877881200151X, doi:https://doi.org/10.1016/j.enbuild.2012.03.003.

[211] M. W. Ulmer, D. C. Mattfeld, and F. Köster. Budgeting time for dynamic vehicle routing with stochastic customer requests. *Transportation Science*, 52(1):20–37, 2018.

[212] M. W. Ulmer, N. Soeffker, and D. C. Mattfeld. Value function approximation for dynamic multi-period vehicle routing. *European Journal of Operational Research*, 269(3):883–899, 2018.

[213] UNAIDS. Malawi: HIV and AIDS estimates. Accessed on 2022-06-21. URL: https://www.unaids.org/en/regionscountries/countries/malawi.

[214] UNAIDS. 90-90-90: An ambitious treatment target to help end the AIDS epidemic, 2014. URL: http://www.unaids.org/sites/default/files/media_asset/90-90-90_en.pdf.

[215] UNAIDS. Malawi 2019 country factsheet, 2019. URL: https://www.unaids.org/en/regionscountries/countries/malawi.

[216] UNAIDS. 2020 global AIDS update - seizing the moment, 2020.

[217] UNITAID. *HIV/AIDS Diagnostics Technology Landscape, 5th edition*. World Health Organization, 2015.

[218] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.

[219] V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. K. Suykens. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011.

[220] P. van der Putten and M. van Someren. CoIL Challenge 2000: The insurance company case, 2000. Accessed on 2019-04-25. URL: https://kdd.ics.uci.edu/databases/tic/tic.data.html.

[221] W. Venter, J. Coleman, V. L. Chan, Z. Shubber, M. Phatsoane, M. Gorgens, L. Stewart-Isherwood, S. Carmona, and N. Fraser-Hurt. Improving linkage to HIV care through mobile phone apps: randomized controlled trial. *JMIR mHealth and uHealth*, 6(7), 2018. doi:10.2196/mhealth.8376.

[222] H. Verhaeghe, S. Nijssen, G. Pesant, C.-G. Quimper, and P. Schaus. Learning optimal decision trees using constraint programming. *Constraints*, pages 1–25, 2020.

[223] P. J. M. Verweij and H. C. Van Houwelingen. Penalized likelihood in Cox regression. *Statistics in Medicine*, 13(23-24):2427–2436, 1994.

[224] S. Verwer and Y. Zhang. Learning optimal classification trees using a binary linear program formulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1625–1632, 2019.

[225] A. Violari, M. F. Cotton, D. M. Gibb, A. G. Babiker, J. Steyn, S. A. Madhi, P. Jean-Philippe, and J. A. McIntyre. Early antiretroviral therapy and mortality among HIV-infected infants. *New England Journal of Medicine*, 359(21):2233–2244, 2008.

[226] F. Vogt, K. Tayler-Smith, A. Bernasconi, E. Makondo, F. Taziwa, B. Moyo, L. Havazvidi, S. Satyanarayana, M. Manzi, M. Khogali, and A. Reid. Access to CD4 testing for rural HIV patients: findings from a cohort study in Zimbabwe. *PLOS ONE*, 10(6):e0129166, 2015.

[227] M. Wen, J.-F. Cordeau, G. Laporte, and J. Larsen. The dynamic multi-period vehicle routing problem. *Computers & Operations Research*, 37(9):1615–1623, 2010.

[228] M. Wen, J. Larsen, J. Clausen, J.-F. Cordeau, and G. Laporte. Vehicle routing with cross-docking. *Journal of the Operational Research Society*, 60(12):1708–1718, 2009.

[229] B. Woods, P. Revill, M. Sculpher, and K. Claxton. Country-level cost-effectiveness thresholds: initial estimates and the need for further research. *Value in Health*, 19(8):929–935, 2016.

[230] World Bank. World Bank Data: Malawi. Accessed 2022-06-21. URL: https://data.worldbank.org/country/malawi.

[231] World Health Organization. Tuberculosis profile: Malawi, 2020. URL: https://worldhealthorg.shinyapps.io/tb_profiles/?_inputs_&lan=%22EN%22&iso2=%22MW%22&main_tabs=%22est_tab%22.

[232] World Health Organization. *Consolidated guidelines on HIV prevention, testing, treatment, service delivery and monitoring: recommendations for a public health approach.* World Health Organization, 2021.

[233] World Health Organization. Guidelines: updated recommendations on HIV prevention, infant diagnosis, antiretroviral initiation and monitoring. 2021.

[234] C. Wright, S. Rupani, K. Nichols, Y. Chandani, and M. Machagge. What should you deliver by unmanned aerial systems? White paper, JSI Research & Training Institute, Inc., and Llamasoft, 2018.

[235] I.-C. Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797–1808, 1998.

[236] I.-C. Yeh. Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, 29(6):474–480, 2007.

[237] I.-C. Yeh and T.-K. Hsu. Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing*, 65, 04 2018. doi:10.1016/j.asoc.2018.01.029.

[238] C. Zhang, A. Atasu, T. Ayer, and L. B. Toktay. Truthful mechanisms for medical surplus product allocation. *Manufacturing & Service Operations Management*, 22(4):735–753, 2020.

[239] F. Zhou, Q. Claire, and R. D. King. Predicting the geographical origin of music. In *2014 IEEE International Conference on Data Mining*, pages 1115–1120, Dec. 2014. doi:10.1109/ICDM.2014.73.

[240] Y. Zhou and J. J. McArdle. Rationale and applications of survival tree and survival ensemble methods. *Psychometrika*, 80(3):811–833, 2015.