Quality Use of Pathology Program (QUPP) - Final Report (July 2020)
Prepared for the Commonwealth Department of Health

Project title:

Exploration of Integrated NATA - RCPAQAP Predictive Modelling to Improve Pathology Quality

Brett A. Lidbury,

Gus Koerbin, Alice M. Richardson and Tony Badrick

Project title: Exploration of Integrated NATA - RCPAQAP Predictive Modelling to Improve Pathology Quality

Project Leader: **Brett A. Lidbury**, Associate Professor, National Centre for Epidemiology and Population Health, The Research School of Population Health, The Australian National University.
Co-Investigators: **Gus Koerbin** (Retired), formerly Chief Scientist - NSW Health Pathology, Chatswood, NSW, **Alice M. Richardson**, Director and Associate Professor, The Statistical Consulting Unit, The Australian National University, and **Tony Badrick**, CEO, Royal College of Pathologists (Quality Assurance Programme), St. Leonards, NSW.

Acknowledgements: The authors wish to thank both pathology (anonymous) networks for their ongoing support of these research investigations via data provision and advice.

Thanks to the staff in the RSPH - Health and Medicine Research Office (ANU) for general administrative support and the facilitation of contracts, payments and agreements between the ANU and the Commonwealth Department of Health.

Objective and Aims (as articulated in the 2017 Project *Funding Agreement*, and subsequent 2020 *Deed of Variation*):

(1) *The **Objective** of the Activity is to provide the Department with a model that will identify poorly performing pathology laboratories across Australia. The model will be developed through the use of established statistical modelling strategies that combine National Association of Testing Authorities (NATA) data plus … quality assurance [programme] (RCPAQAP) data to explore the challenges to achieve consistent, high-quality pathology laboratory performance*, and **(2)** *The **Aim** of the Activity is to proactively detect and resolve poor laboratory performance using sophisticated computational knowledge-discovery methods (primarily machine learning), and relationships between two distinct sets of performance metrics (NATA and RCPAQAP)*.

For noting - The above-stated Aims/Objectives are not concerned with identifying individual laboratories, but to develop system-level models to allow the early detection of quality-control challenges. By extension, to assist in the remediation of quality control issues as soon as possible, to ensure the best quality pathology results and decision-making by scientists and pathologists for the clinicians they support via laboratory diagnosis and monitoring.

The data provided for this research project were obtained directly from the Pathology networks studied, with no involvement from the RCPAQAP or NATA.

Executive Summary:

As summarised in the original and updated funding agreements, all stated aims/objectives, as represented by the B.3 indicators/targets, were addressed and achieved. The primary thrust of the research investigations dealt with NATA and RCPAQAP results obtained from a State government pathology network (State Pathology Laboratory Network - SPLN), and a private laboratory with state-wide coverage (Private Pathology Laboratory Network). The objective of integrating NATA and RCPAQAP results was achieved, with the subsequent aim of providing a model as a result of NATA/RCPAQAP integration. Of note however, was the very different NATA profiles for SPLN versus the PPLN, which made direct comparisons difficult, but lead to a useful results nonetheless. As found via an earlier pilot study, bias calculated from GGT and serum creatinine RCPAQAP results remained as leading predictors of NATA performance for a government (state) pathology network. The reason for the enhanced utility of GGT (RCPAQAP) bias as a leading predictor of NATA was determined by subsequent investigations of distribution around the specific RCPAQAP target value.

A range of issues, problems or delays were encountered, including, as mentioned above, disruption due to the bushfire crisis of late 2019 - early 2020. In terms of meeting agreed milestones, the delay in receiving point-of-care-testing (PoCT) data had the most significant impact on milestone compliance (hence the March - April 2020 *Deed of Variation*). The PoCT data received covered three geographical zones under SPLN administration; therefore, while differences could be examined in this context, the proposed analysis of different states and pathology providers could not be fulfilled. Also, the ambiguities in the raw data received, in general, resulted in the data-cleaning phase of the project proceeding slower than anticipated. Careful data cleaning and organisation is critical to the project success, since faulty data sets lead to wrong conclusions.

The key findings from this research investigation were in the development of simple decision rubrics from integrated NATA and RCPAQAP results to support assessments of laboratory quality. The benefit to stakeholders is the provision of relatively simple tools, based on common laboratory RCPAQAP markers with which to assess laboratory performance. The distribution of marker (bias) results over the sixteen-point RCPAQAP cycle, and decision boundaries discovered by machine learning, are the bases of the decision support strategies presented.

In spite of these unforeseen challenges, the aims were mostly addressed (Appendix A), with results and conclusions delivered in line with the original research proposal. As alluded to above, some variations to the original proposed aims may be found due to unforeseen issues

with data scope or access, but in general, results were forthcoming that addressed the need for an integrated NATA - RCPAQAP model to efficiently detect quality control challenges.

Glossary:

Alb - Serum albumin
ALT - Alanine Aminotransferase
AST - Aspartate Aminotransferase
Bicarb. - Sodium bi-carbonate
Category – Class *
CV% - Coefficient of Variation percent
Creat. - Serum creatinine
GGT - Gamma (γ) Glutamyl-Transferase
LD - Lactate dehydrogenase
LFT – Liver Function Test
NATA - National Association of Testing Authorities
PLS - Partial Least Squares
PPLN - Private Pathology Laboratory Network
pCO_2 - Blood gas CO_2 analysis
PoCT - Point-of-Care Testing
SPLN - State Pathology Laboratory Network
RDW – Red cell Distribution Width
RF(A) – Random Forest (Analysis)
RCPAQAP - Royal College of Pathologists of Australasia Quality Assurance Programme
SD - Standard Deviation
SVM – Support Vector Machine
Trees - single decision trees (recursive partitioning) → Forests)
TBil - Total serum bilirubin (DBil - Direct bilirubin)
TnI - Troponin I
TP, FN - (True Positive, False Negative, TN, FP)
UEC - Urea, Creatinine, Electrolytes

** Class/Category are used interchangeably*

Period of activity:
Total Project (November 2017 - June 2020); Final Report (July/August 2019 – July 2020).

Scope of Work:
This document will report on the aims, subsequent research activities and results, as stated in the original *Funding Agreement* (2017), and updated *Deed of Variation* agreement finalised during March - April 2020 (project extension due to bushfire impacts and the late receipt of PoCT data for analysis).

As stated in the above agreements, the following activities were conducted and project aims achieved:

1. *Compile and clean NATA/RCPAQAP datasets - **(a)** Liaise with a State Pathology Laboratory Network (SPLN), a Private Pathology Laboratory Network (PPLN), and the Royal College of Pathologists of Australasia Quality Assurance Programs Pty Ltd (RCPAQAP), to access past NATA audit and external quality assurance/quality assurance programme (EQA/RCPAQAP) assay performance data. Thereafter - **(b)** Extract raw data, assess and correct for missing values, censored and incomplete data and undertake initial statistical analyses; and **(c)** Ensure modelling of these data allows for overlapping markers of poor pathology laboratory performance and the development of performance metrics that will predict poor laboratory performance earlier than possible using separate quality control schemes.*

2. *Modelling with decision tree + SVM pattern recognition algorithms - **(a)** Undertake intensive pattern recognition analyses on the combined NATA + RCPAQAP (clean) data set; and **(b)** Produce data pattern "rules" reflected by pathology predictor variables used for quality evaluation, in response to quality outcome variables. Apply to data, and also troponin turn-around-time (TAT) studies.*
3. *Application of results (from 1 and 2 above) to Point-of-Care tests (PoCT) - **(a)** Repeat steps above for the PoCT data set obtained via the SPLN; **(b)** Evaluate the PoCT network and determine if there are associations between NATA assessment and RCPAQAP data (Creatinine, Troponin, and INR); **(c)** Using a small subset of the PoCT sites, use the NPAAC - PoCT Guideline as an assessment guide to assess whether there are associations between failed Clauses, QC or EQA and performance; **(d)** Obtain NATA and RCPAQAP data from another network (PPLN) and validate the findings using a different geographical patient base for validation; and, **(e)** Develop 'rules' for the prediction of PoCT quality, as conducted via identical statistics and machine learning methods.*
4. *"How the results of this Activity will be used to benefit pathology stakeholders"* - for commentary on this aspect, please see the report Discussion.

Context: Linked RCPAQAP results and NATA reports were obtained from a large, state-wide private pathology enterprise (PPLN), and a state-wide government pathology network (SPLN). PoCT data were only received for three regions under the state-wide government network. The profile of NATA reports, as reflected by the number and variety of major and minor conditions, and/or observations, were different for the private and state examples.

Changes, limitations experienced during the QUPP funding period:

During the project period, unexpected challenges were encountered that impacted originally proposed timelines and project details. Examples include -

- The retirement of Dr Gus Koerbin from his position in the time leading up to the project start date. This impacted data access in the short-to-medium term;
- As described above, and linked to the point above, access to point-of-care testing (PoCT) data was slow, and not delivered for analysis until early 2020;
- Also, in terms of PoCT, data were only provided by SPLN (for three regions), so a wider quality performance comparison with other networks was not possible at this time;
- Impact of bushfires and associated evacuations on work flow over early 2020 (workflows from March onwards were also impacted by the COVID-19 lockdown).

(1) Introduction

In Australia, pathology laboratory performance is evaluated via conformance to ISO 15189 and National Pathology Accreditation Advisory Council (NPAAC) standards, and external quality assurance (EQA) processes (also known as “RCPAQAP” - Quality Assurance Programme), established to ensure ongoing excellence in pathology laboratory quality. The assessment of ongoing ISO 15189 and NPAAC standards is the responsibility of the National Association of Testing Authorities (NATA) (1), which uses regular laboratory audits and the reporting of non-conformances (e.g. Conditions - immediate action required) to ensure the maintenance of the highest analytical, management and technical standards behind the provision of pathology results, which are recognised as a central pillar to modern medical practice. The *Royal College of Pathologists of Australasia* (RCPA) QAP process is an essential arm of the maintenance of these necessary high standards, via the direct assessment of NATA-accredited laboratories capacity to accurately measure the range of routine and special test markers from patient samples. These two distinct processes are connected in the pursuit of excellence for pathology performance, but traditionally have not been integrated into a single model of quality assessment, or unified system of monitoring and advice.

The report authors have previously completed a QUPP pilot study into the potential of enhancing the overall quality endeavour via the integration of NATA and RCPAQAP (EQA) results and data (2), which was subsequently published in the *Journal of Laboratory and Precision Medicine* (3). The impetus for this research enquiry into quality control in pathology testing was emphasised further by the accompanying systematic scoping review, conducted as part of the wider investigation, which found no sophisticated methods in the existing peer-reviewed literature for the monitoring and remediation of quality issues in laboratories beyond standard “root cause analysis” (2, 3). Recognising this gap, data modelling via machine learning algorithms and supporting statistical tests were conducted on NATA and RCPAQAP data provided by a participating state-wide, government pathology network.

This project report extends the data modelling component of the original pilot study. With the aim of validating pilot study results, data from another State Pathology Laboratory Network (SPLN) RCPAQAP - NATA cycle were explored, as well as RCPAQAP - NATA results obtained from a private pathology network (PPLN) with state-wide coverage. With this opportunity, similar analyses were also conducted on Point-of-Care Testing (PoCT) RCPAQAP - NATA results, provided by the participating SPLN.

(2) Methods

The analyses, observations and conclusions presented by this report were the result of a variety of statistical methods of data interrogation, as well as predictive modelling via machine learning (ML) algorithms. The ML methods applied (and described in detail below)

were two recursive partitioning algorithms (decision trees and random forest), as well as support vector machines (SVM). All methods have analytical advantages and disadvantages, but when used in tandem provide an excellent predictive modelling platform, from the broader patterns of interaction between response and explanatory variables, to the specific decision thresholds discovered that support outcome prediction.

(a) Statistical Methods

Standard statistical analyses, namely the calculation of sample mean, median, standard deviation, histograms and other plots, the Runs test, One-Sample Wilcoxon Signed Rank Test and ANCOVA, were performed using SPSS for Windows (version 26) (4).

A p-value of 0.05 was applied to decide statistical significance.

(b) Calculation of Bias for RCPAQAP Markers

The RCPAQAP results were calculated from individual bias scores across the 16 time points of the RCPAQAP cycle (11 time points for PoCT RCPAQAP), thus representing how close the laboratories were able to achieve the RCPAQAP target value for the specific test time point (1 - 16), which was determined from the RCPAQAP mean calculated from the national data comprising 450 - 600 laboratories (RCPAQAP target values for PoCT were calculated from 140 - 170 laboratories).

The bias results were presented as the raw result, and not a percentage, where a bias outcome of zero (0) was a perfect RCPAQAP result (namely, has achieved the exact RCPAQAP target value), or can be represented as a negative (-) or positive (+) result, indicating a laboratory RCPAQAP result below or above the RCPAQAP target value, respectively, with the degree of variation from zero on a continuous scale, indicating the extent of variation from the target value.

The equation to calculate bias from RCPAQAP UEC, LFT (etc) results is -

$$\frac{\textbf{(Lab result - RCPAQAP Target value)}}{\textbf{RCPAQAP Target value}}$$

For Example -

ALT Result (RCPAQAP Time point 5, Labs 1 and 15)

$$\frac{(55 - 51)}{51}$$

= (+) 0.078 Bias **(Lab 1)**

$$\frac{(47 - 51)}{51}$$

= (-) 0.078 Bias **(Lab 15)**

It is essential to note that unless stated otherwise, all RCPAQAP data entered into machine learning models, statistical analyses and data plots, were first transformed into relative bias values, as achieved via the above equation.

(c) Machine Learning - Integration of NATA and RCPAQAP results

The R Statistical programming language was used for all machine learning analyses, and ultimate construction of NATA - RCPAQAP prediction models (5). The R package e1071 was used for support vector machine (SVM) modelling (6, 7). For recursive partitioning, two algorithms were employed; (i) the rpart package for single decision trees (8), and (ii) the randomForest (RF) package for analyses of the same name (9).

In tandem with the above R packages, the caret package was used extensively for the tuning and inspection of machine learning models, particularly for random forests (10). This package supported the tuning of RF models by identifying optimal "mtry" parameters (i.e. number of trees per decision node), and by receiver operating characteristics (ROC). As well as model fitting, caret assessed model accuracy and robustness. For RF, the kappa and McNemar's statistics were generated to allow the assessment of model efficiency.

The models developed for all laboratories and PoCT involved the prediction of NATA Class (category) by multiple RCPAQAP bias results. Therefore, the NATA Class acted in the model as the response to be predicted by RCPAQAP bias (explanatory or predictor) variables.

(d) Attribution of NATA Results into Response Classes

The results of NATA inspections utilised in this study were presented as written feedback accompanied by recommendation classes - namely a (i) *Condition*, (ii) *Minor* (condition) and/or (iii) an *Observation*. Throughout the report these are often represented a simply "C", "M" or "O". A NATA Condition is a recommendation that requires immediate attention and resolution before the following inspection. If not addressed adequately, Conditions can lead to laboratory closure or other sanction. Minor (conditions) require attention, which if not addressed, risk becoming a full Condition during the following inspection.

Both C and M reports/recommendations are taken as the primary responses of interest for modelling, since they represent technical and/or management problems at the laboratories involved, and hence may reflect in the RCPAQAP results as difficulties in obtaining accuracy (i.e. achieving the RCPAQAP designated target value, or within an acceptable error range). By broad definition, the NATA - RCPAQAP (Bias) modelling is attempting to link systemic operational problems detected by NATA with RCPAQAP performance. The value here is the potential that RCPAQAP metrics/models may assist to

identify system-level issues in advance, resulting in fewer C and M reports on subsequent NATA inspection.

As well as C and M reports, NATA also records “Observations” (O), which are not necessarily critical or require action, but are provided by assessors to assist the laboratory in maintaining standards. O reports were not used if sufficient C and M reports were available.

(i) Private Pathology Laboratory Network - NATA Classes for the PPLN were “Yes” or “No”. As reported previously (Performance Report 3; Table 1) around 50% of the PPLN laboratories in the sample did not record C, M or O reports, leading to the designation of these laboratories to the No NATA Class. Alternatively, the presence of a C, M or O resulted in the inclusion of the responsible laboratory in the Yes NATA Class.

(ii) State Pathology Laboratory Network - As reported in the pilot study from 2107 (2), and found again for this project (Table 2), SPLN laboratories attracted many C and M reports, and hundreds of Observations. Therefore, to create NATA Classes for RCPAQAP interrogation, only C and M were counted. The NATA Classes for SPLN, therefore, were “High” or “Low”. Whether a laboratory was classified as high or low was decided by the incidence of C or M reports as above or below the median calculated for all laboratories. The low NATA Class contained laboratories with zero reports, but these were a minority.

(iii) Point of Care Testing - NATA and RCPAQAP results were available for three SPLN regions, with the NATA reports generalised to all of the laboratories within the specific region. On inspection, there were not large differences in the number of NATA C, M or O reports between the three regions, so while analysis could be performed, the subsequent results were not indicative of a link to NATA performance.

(3) Results - Project Activity Reports

(a) NATA Profiles

(i) Private Pathology Laboratory Network (PPLN) – The PPLN sample comprised 22 B laboratories, with only 2 G laboratories available for investigation. Therefore, a comparison of B and G laboratories was not possible due to inadequate G laboratory sample size.

Table 1 summarises the profile of NATA results for the sample of 23 - 24 private laboratories included in this study, supported by Appendix B (and first reported in the January 2019 Performance Report). A total of forty-eight (48) Observations (O), Minor conditions (M) and major Conditions (C) were reported, and of these, only 2 were major conditions requiring urgent attention. Unlike the SPLN NATA results (below), observations (O) were included in the NATA profile to ensure sufficient scope to develop response categories for analysis against RCPAQAP results. Observations were generally advice of a

non-urgent nature, so could not be assumed to concern a laboratory breach (see Appendix C for the range of NATA comments and advice received).

The general modelling of PPLN NATA-RCPAQAP results, therefore, used total NATA CMO counts to develop response classes (absence or presence of NATA reports), although more sophisticated (and accurate) SVM models were ultimately developed from counting only the number of M reports (see Results).

The NATA feedback and further details on these results can be found in Appendices B and C of this report, as well as referring to Performance Report 3 from January 2019.

TABLE 1 – Results from NATA inspections conducted on the Private Pathology Lab Network (PPLN) included in this study. Only laboratories with a minimum of 1 O, M or C are included. Full NATA details for the PPLN inspected are available in Appendix B to this report. Laboratories were not divided into B or G categories. *(Presented in Performance Report 3 - January 2019).*

NATA Details	Conditions (C)	Minor Report (M)	Observation (O)
Number recorded	2	36	10
Associated NATA Clause(s)	4.13, 5.1	4.1, 4.10, 4.13, 4.14, 4.3 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.8	4.1, 4.10 5.1, 5.2, 5.3, 5.4, 5.5, 5.6
Example Comments from NATA Reports	▪ *Trauma pack fate – (no documents)* ▪ *Supervising scientist working hours*	▪ *Relief … supervising scientist* ▪ *Hardcopy documents in laboratory* ▪ *Obsolete versions of documentation* ▪ *1-2 weekly, maintenance incomplete* ▪ *Continuing education program all staff* ▪ *Drift fail not investigated/actioned*	▪ *Improve root cause/corrective action* ▪ *Competence not accessed on day* ▪ *Collection staff - incorrect draw order*

NATA – National Authority of Testing Authorities See Appendix B for full NATA details.

(ii) State Pathology Laboratory Network (SPLN) - The NATA results recorded had many differences in comparison to the PPLN (Table 2: Appendix D). These were: (1) a larger sample size (41 laboratories in total); (2) a sufficient number of laboratories to allow comparisons between B and G designated laboratories; and (3) around 10-fold more total CMO reports, with O reports greater than 100, and C and M reports sufficient to develop models without O data (that were too abundant and diverse to capture meaningful trends.

Further investigations may consider deeper analyses of Observation feedback as a guide to performance, but in this instance, O reports generally were not critical).

Given these differences in NATA results, deeper analyses were possible for SPLN (Table 2). The mean ($\bar{x}$) Conditions were 3.38 per laboratory in the sample investigated, with Minor Reports significantly higher ($p < 0.001$) at 6.26 per laboratory. The separation of C and M reports according to the laboratory category, B or G, revealed differences due to these designated laboratory roles within the SPLN system. For C reports, a significantly higher mean was found for G laboratories compared to B laboratories ($p < 0.005$), with a significantly higher mean also found for M reports associated with G laboratories ($p < 0.005$) (Table 2).

With the B laboratories supervised by larger G laboratories, it is likely that the significantly higher number of C and M NATA reports reflect the comparative size of the laboratories, and range of testing conducted. The likelihood is also that the responses to NATA reporting from the G laboratories was transferred to the B laboratories that they supervise. The relationship between G and B laboratories in the context of NATA, RCPAQAP and laboratory quality modelling requires a larger dataset, and a dedicated focus.

TABLE 2 – Results from NATA inspections conducted on the SPLN laboratories (n = 41) included in this study. The total cohort comprised B (n = 31) and G (n = 10) category laboratories (Appendix D). *(Presented in Performance Report 4 - July/August 2019)*

NATA Details	Conditions (C)		Minor Report (M)		Observation (O)
Number recorded	128		244		> 100
Mean ($\bar{x}$) Reports/ Laboratory	3.28		6.26 *		NA
Mean ($\bar{x}$) Reports B and G Laboratories	B	G	B	G	NA
	2.4	5.4 *	5.0	8.8 *	NA
Associated NATA Clause(s)	4.2, 4.9, 5.1, 5.4, 5.5.		4.1, 4.3, 5.1, 5.3, 5.5, 5.6.		Across most categories – not critical reports
Example Comments from NATA Reports	▪ *Quality management across all departments* ▪ *Investigation + corrective action – root cause* ▪ *IATA training required + other training inadequacies* ▪ *AS/NZS 4308 adherence (tampering)*		▪ *Records maintenance* ▪ *Documents and worksheets missing, not updated* ▪ *Corrective action requests (CARs) – timeframes* ▪ *POC training manual* ▪ *Validation – non standard methods*		*Extensive comments and positive or helpful feedback across most NPAAC Clauses*

NATA Details	Conditions (C)	Minor Report (M)	Observation (O)
	▪ *Procedures for Westgard Rules and evaluation* ▪ *TGA notifications (IVD framework)*	▪ *Staff engagement with RCPAQAP material*	

NATA – National Authority of Testing Authorities SPLN – State Pathology Laboratory Network

* Significantly increased mean for M versus C NATA reports, and the same reports for B versus G laboratories (independent T-test, 2-tails, equal variance. Significance $p < 0.001$ for all comparisons).

(iii) PPLN versus SPLN (Conclusion) - The profile of NATA results, as captured via the reporting of major Conditions (C), Minor conditions (M) and/or Observations (O), was bimodal in nature, with PPLN reporting very few conditions (e.g. only 2 x C records from 24 laboratories - Table 1), while SPLN recorded hundreds of C and M reports (Table 2) that were still exceeded after the PPLN laboratory sample size was weighted to align with the larger SPLN laboratory number. The aim of this project is not to explain the differences in NATA results between the two pathology networks, but to integrate the NATA results, as reflected by conditions and/or observations, with RCPAQAP cycles from the same time period. What is relevant, however, is that NATA profiles for the PPLN and SPLN do not allow a direct comparison between the state - private systems, or validation across geographical boundaries. What is offered, alternatively, are rules for systems that have different NATA experiences.

(b) RCPAQAP Performance and Relationships with NATA Results

A range of RCPAQAP results were analysed against percentile rankings, coefficient of variation (%), and other descriptive measures of data dispersion for each laboratory in the sample, prior to machine learning. Rankings based on these measures were aligned to the number of NATA C and M reports to assess broadly whether RCPAQAP rankings agreed with NATA - assessed performance.

Data were available for sixteen (16) time points over a RCPAQAP cycle, and represented a range of routine and special (e.g. drugs, antibiotics) blood or serum tests. The analyses conducted from here onwards focussed on routine urea, creatinine, electrolyte (UEC) RCPAQAP evaluations, and liver function test (LFT) markers. The UEC and LFT markers were available from both the PPLN and SPLN data, and were available for the majority of laboratories included for the PPLN and SPLN data sets (special test RCPAQAP was provided for a minority of laboratories from the SPLN sample, with no special test RCPAQAP data available from PPLN).

(i) PPLN Laboratories - Table 3 summarises the ranking results for PPLN laboratories, focussed on three RCPAQAP markers that were previously found to be effective predictors

of NATA outcomes: GGT, serum creatinine and serum potassium (2). The laboratories were ranked in descending order by the number of NATA M observations recorded individually, with corresponding standard deviation (SD), percentile, coefficient of variation percent (CV%) and specific laboratory bias matched with the NATA M rankings.

The hypothesis for consideration is that the individual laboratories with zero NATA minor (M) reports will have the highest RCPAQAP percentile rankings (e.g. < 20%) for the three markers investigated, while the laboratories that recorded NATA minor (M) conditions (1 - 7 for the laboratories represented) would show poorer performance as assessed by their RCPAQAP results, with the higher number of M conditions associated with poorest performance.

TABLE 3 - Summary of laboratory rankings for **(a) serum GGT**, **(b) serum Creatinine,** and **(c) serum Potassium.** Quality assurance programme (RCPAQAP) compared to the number and type of NATA observations, for a sample of individual laboratories from Private Pathology Laboratory Network. Laboratories were ranked in descending order by the number of Minor (M) NATA observations.

(a) GGT

Lab	S.D.	Percentile	CV%	Bias	NATA Observations			
					Condition	Minor	Observe	NATA Total
16	2.3	61	3	9.4	2	7	1	10
20	1.5	22	1.9	10.4	0	7	0	7
15	2.6	69	3.2	9.3	0	6	3	9
21	1.6	28	2	9	0	4	3	7
14					0	3	0	3
13	1.9	45	2.5	10.9	0	2	0	2
19	1.5	24	1.9	10.5	0	2	1	3
6	1.5	20	2	9.8	0	1	0	1
11	2.5	68	3.1	8.9	0	1	0	1
1	0.8	1	1.1	9.6	0	0	0	0
2	1.9	46	2.4	9.6	0	0	0	0
3	1.4	17	1.8	8.8	0	0	0	0
4	2.1	56	2.7	11.4	0	0	0	0
5	2.9	76	3.7	11.1	0	0	0	0
7	1.7	35	2	7.7	0	0	0	0
8	1.7	34	2.2	11.1	0	0	0	0

Lab	S.D.	Percentile	CV%	Bias	NATA Observations			
					Condition	Minor	Observe	NATA Total
9	2.1	56	2.7	8.7	0	0	1	1
10	1.3	12	1.6	9.8	0	0	0	0
12	2.1	55	2.7	10.2	0	0	0	0
17	1.3	12	1.6	9.8	0	0	0	0
18	2.2	59	2.8	10.5	0	0	0	0
22	1.7	34	2.3	12.4	0	0	0	0

(b) Creatinine

Lab	S.D.	Percentile	CV%	Bias	NATA Observations			
					Condition	Minor	Observe	NATA Total
16	7.8	88	3.9	8.5	2	7	1	10
20	3.1	7	1.6	11.7	0	7	0	7
15	4.1	27	2.1	9.8	0	6	3	9
21	4.7	38	2.4	8.4	0	4	3	7
14					0	3	0	3
19	5.6	58	2.8	10.3	0	2	1	3
13	4.9	41	2.5	12.2	0	2	0	2
6	3.7	18	1.9	10	0	1	0	1
11	7.2	83	3.6	6.3	0	1	0	1
9	4.4	32	2.2	7	0	0	1	1
1	7.7	87	3.8	13.3	0	0	0	0
2	7.1	81	3.6	11.6	0	0	0	0
3	3.8	20	1.9	5.9	0	0	0	0
4	5.7	61	2.9	9.4	0	0	0	0
5	4.6	37	2.3	9.9	0	0	0	0
7	5.8	64	2.9	6.7	0	0	0	0
8	4.1	27	2.1	8.8	0	0	0	0
10	2.7	4	1.4	5.4	0	0	0	0
12	8	89	4	10.8	0	0	0	0

Lab	S.D.	Percentile	CV%	Bias	NATA Observations			
					Condition	Minor	Observe	NATA Total
17	6.1	70	3.1	9.8	0	0	0	0
18	6.9	79	3.5	11.8	0	0	0	0
22	4.7	38	2.2	1.4	0	0	0	0
24	5.2	49	2.7	12.8	0	0	0	0

(c) Potassium (K^+)

Lab	S.D.	Percentile	CV%	Bias	NATA Observations			
					Condition	Minor	Observe	NATA Total
16	0.09	82	2.1	0.02	2	7	1	10
20	0.09	80	2.2	0.06	0	7	0	7
15	0.05	5	1.2	0.01	0	6	3	9
21	0.08	71	1.9	0.03	0	4	3	7
14					0	3	0	3
13	0.08	73	1.9	0.03	0	2	0	2
19	0.07	52	1.6	0.02	0	2	1	3
6	0.06	31	1.5	0.04	0	1	0	1
11	0.07	60	1.8	0.03	0	1	0	1
1	0.06	19	1.4	0.03	0	0	0	0
2	0.07	50	1.6	0.02	0	0	0	0
3	0.08	67	1.9	0.06	0	0	0	0
4	0.06	35	1.5	0.04	0	0	0	0
5	0.1	90	2.3	0.07	0	0	0	0
7	0.08	75	2	0.01	0	0	0	0
8	0.07	53	1.6	0.02	0	0	0	0
9	0.06	32	1.4	0.04	0	0	1	1
10	0.06	28	1.5	0.03	0	0	0	0
12	0.07	42	1.8	0.09	0	0	0	0
17	0.11	94	2.5	0.04	0	0	0	0
18	0.09	84	2.2	0.06	0	0	0	0
22	0.11	94	2.6	0.01	0	0	0	0

S.D. (Standard Deviation) CV% (Coefficient of Variation %).

NATA Observations (from laboratory inspections) – "Condition" (must be addressed as a condition of continuing operation), "Minor" (satisfactorily addressed before the next inspection), "Observe" (Observation by the NATA assessors of interest to the laboratory, to assist the lab's quality regime).

*Laboratories are ranked in descending order by the number of **Minor** NATA observations (beige column). The top ranked laboratory, as decided by percentile, is highlighted in the beige-shaded row.*

For GGT (Table 3a) and serum creatinine (Table 3b) the top ranked laboratories by percentile (lab 1 for GGT, lab 10 for creatinine - highlighted by shading in the tables) both recorded zero NATA M observations, which suggests that the hypothesis on the link between superior NATA and RCPAQAP performances is correct. On further inspection of the GGT results, however, we find that a laboratory (lab 5) with a RCPAQAP percentile of 76 also recorded zero NATA M observations, and conversely, laboratory 20 recorded 7 M observations, with a RCPAQAP percentile of 22.

Laboratory 20 also recorded a higher percentile ranking of 7 for creatinine, in spite of the 7 M reports/observations. For GGT and creatinine RCPAQAP results (Tables 3a & b), in general, no link between NATA and RCPAQAP performance was observed when evaluating individual laboratories. This was confirmed by bivariate correlation analyses, which showed no significant associations between RCPAQAP percentile rankings and the number of NATA M observations reported (results not shown). This was true also for serum potassium.

For serum potassium (Table 3c) the top-ranked laboratory in relation to RCPAQAP percentile ranking was lab 15, but which also received 6 M observations from NATA, with only two laboratories attracting more M observations (7), indicating a mis-match between NATA and RCPAQAP quality ratings. Conversely, the lab with the most M observations (in addition to 2 x C reports and 1 x O), and the most total NATA reports of all laboratories investigated (lab 16), was ranked at a RCPAQAP percentile of 82, which suggests a link between NATA and RCPAQAP quality evaluations. To further emphasise the inconsistencies between NATA reports and RCPAQAP results, of the laboratories with zero NATA C, M or O (observations), three had percentile rankings above 90% - namely laboratories 5, 17 and 22.

Certain laboratories in isolation demonstrated the hypothesised relationship of zero (or low) NATA reports/observations with high RCPAQAP percentile rankings, suggesting that the lab performance as evaluated by either NATA or RCPAQAP is associated, and reflects laboratory quality impacts broadly. However, this was not a pattern found for the PPLN laboratories when considered collectively.

(ii) SPLN Laboratories - As described above, the SPLN laboratory sample was larger and comprised a majority of B category laboratories, with around the 25% of the sample being supervising G category laboratories. Also notable was the large number of NATA reports recorded for SPLN laboratories (Table 2). With these differences, direct comparison with PPLN NATA and RCPAQAP performance was not possible, but some trends remained consistent.

With the larger number of C and M observations, the comparison of RCPAQAP percentiles (represented for analysis by quartile ranges) was achieved by calculating NATA C, M and C+M group means ($\bar{x}$ ± SEM) and presenting them alongside the RCPAQAP

quartile rankings (1 represents the best performing laboratories with percentiles of ≤ 25%, with quartile 4 the worst performers with percentiles of ≥ 75%).

The expectation was that with a decrease in RCPAQAP quartile ranking, the mean number of NATA C, M and C+M reports/observations would increase, reflecting a trend from excellent to poor laboratory performance. Like for the PPLN results, this was not found, with the results (Table 4) being random; for example, the NATA C - M means being similar or higher for RCPAQAP quartile 1 when compared to quartile 4. This was a feature of all three RCPAQAP markers examined (GGT, creatinine, potassium).

(iii) Conclusion - While differences in NATA profile are present when comparing PPLN to SPLN, the two laboratory networks share a common feature in that the results of the NATA inspections do not reflect the results of the RCPAQAP cycle, in this case, represented by previously identified serum analytes with the best predictions of NATA categories, namely GGT, creatinine and potassium (2).

TABLE 4 - RCPAQAP quartile (%) rankings for SPLN laboratories according to: **(a)** serum Creatinine, **(b)** GGT, and **(c)** serum Potassium, and matched with the Mean ($\bar{x}$) NATA conditions for each quartile (descending from the highest RCPAQAP performance to the lowest).

RCPAQAP Marker	Percent (%) Quartile	NATA Conditions Quartile Means ($\bar{x}$) (± SEM)		
		Minor (M)	Major (C)	Minor + Major
Serum Creatinine	1	3.6 ± 0.9	2.0 ± 1.0	5.6 ± 1.7
	2	6.4 ± 0.9	3.8 ± 0.6	10.1 ± 1.4
	3	7.7 ± 1.4	4.2 ± 1.3	11.9 ± 2.6
	4	4.8 ± 1.0	1.0 ± 0.6	5.8 ± 1.3
GGT	1	6.6 ± 1.0	5.1 ± 0.9	11.6 ± 1.9
	2	6.3 ± 1.4	3.0 ± 1.0	9.3 ± 2.2
	3	4.0 ± 0.9	1.7 ± 0.5	5.7 ± 1.2
	4	7.1 ± 1.4	2.5 ± 1.3	9.6 ± 2.5
Serum Potassium	1	6.3 ± 1.0	3.5 ± 0.9	9.8 ± 1.8
	2	3.8 ± 0.9	1.9 ± 1.0	5.7 ± 1.8
	3	7.3 ± 1.7	4.3 ± 0.9	11.5 ± 2.3
	4	6.4 ± 1.2	3.0 ± 1.0	9.4 ± 2.2

Percent Quartile Ranges: 1 (1 - 25%), 2 (26 - 50%), 3 (51 - 75%), 4 (76 - 100%)
Quartile range 1 = Best RCPAQAP Performance; Quartile range 4 = Worst RCPAQAP Performance.
M - Minor condition reported by NATA. C - major Condition reported by NATA.

Each PPLN and SPLN laboratory in the sample included percentile rankings, CV%, SD and bias specific to individual laboratories across the RCPAQAP cycle of 16 timepoints, and thus providing ranking of each laboratory in relation to RCPAQAP performance.

Laboratory-specific percentiles, which accorded with CV%, were used to rank the laboratories in order from best to worst RCPAQAP performance (Tables 3 and 4). To these rankings the corresponding NATA report data were added for each laboratory, which found no relationship between RCPAQAP ranking and the number of C, M and/or O reports for PPLN laboratories (Table 3), or C, M or C+M total NATA reports for SPLN.

In summary, the matching of RCPAQAP and NATA results does not allow the direct integration of the respective data into a single quality model, suggesting more sophisticated treatment of the data to reveal underlying patterns that link these two data sources.

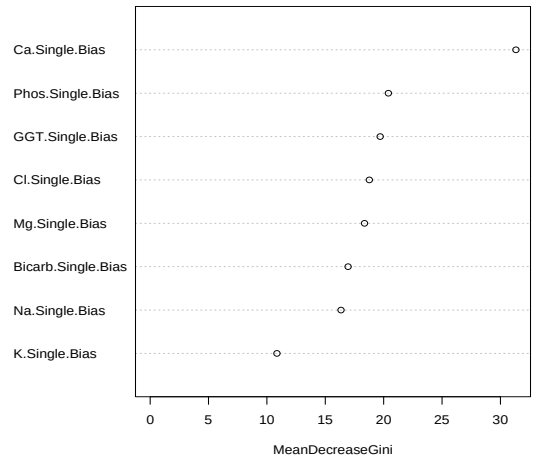
(c) Machine Learning - Prediction of NATA outcomes by RCPAQAP Bias

(i) PPLN Electrolytes - The RCPAQAP data for 22 - 23 anonymous PPLN laboratories all included a range of UEC serum markers, namely - bicarbonate, calcium, chloride, magnesium, phosphate, potassium and sodium, with GGT added as a control RCPAQAP marker (based on observations from our previous pilot study) (2). To interrogate the link between RCPAQAP and NATA reports, RCPAQAP bias values aggregated across all laboratories were used to explain the NATA report categories, representing PPLN laboratories with zero (0) NATA reports, or 1 or more ($\geq$ 1) CMO conditions/observations reported during NATA inspections (NATA Class Yes or No).

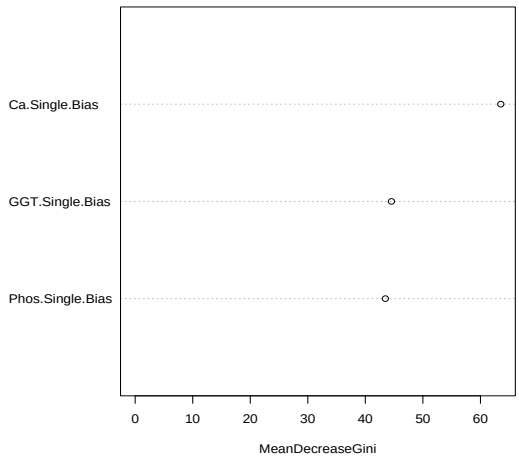
Recursive Partitioning - Using the R packages (rpart) and (randomForest), rankings and RCPAQAP bias thresholds were determined in relation to NATA categories, via recursive partitioning algorithms (Figure 1). Supporting Figure 1 is Table 5, which summarises the performance of the random forest analyses, as well as provides model performance parameters from partial least squares (PLS) investigations (R package - caret). The performance parameters include calculation of positive and negative predictive values.

Results: The results presented in Table 5 suggest that a model utilising only the top three predictors of NATA class (0 or 1) (Figure 1b) was of similar predictive power as the full PPLN electrolyte model (Figure 1a), with serum calcium and phosphate being the best electrolyte predictors, in concert with serum GGT. Final model accuracy was only 2.2% lower for the top three predictors in comparison with the full (8 predictor variable) model, which was reflected also by the Out-of-Bag (OOB) error rate. Positive and negative predictive values (PPV, NPV) were also similar, and for both models specificity was superior to sensitivity, indicating that the recursive partitioning models were more effective at predicting true negative results; however, this was not supported by the ultimate PPV and NPV results, with the opposite suggested.

(a)



(b)



(c)

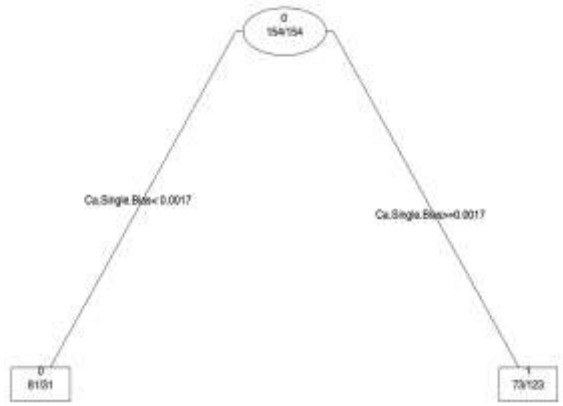


Figure 1 - PPLN RCPAQAP electrolyte bias predictions of NATA class by recursive partitioning - **(a)** Full Random Forest Analysis (RFA) of RCPAQAP electrolyte and GGT bias; **(b)** RFA analysis of the top three predictors identified by RFA model in (a); **(c)** Single decision tree model of the same RCPAQAP electrolyte - GGT bias results to predict NATA CMO classes.

** NATA classes decided as the laboratories that recorded NATA C, M and/or O reports (1), compared to laboratories with no NATA reports recorded (0), which coincided with the RCPAQAP cycle investigated (C - major condition; M - minor condition; O - Observation).*

While final model accuracies and attendant (OOB) error rates were respectable at 66 - 68.5% and 32 - 35% respectively, the measures of model performance were poor, which may explain the inconsistency of the sensitivity and specificity results, in spite of the RF algorithm being tuned prior to analysis for optimal performance. The two statistics to observe in this context are Kappa and McNemar's test results. The Kappa statistics for the final models of 0.33 - 0.37 indicate that the agreement of "votes" for predicting the correct NATA class (0 or 1) were only correct in 33 - 37% of cases. The McNemar's test results for both final models were $p < 0.05$, indicating an imbalance of marginal values in the 2 x 2 confusion matrix, indicating a poor model (McNemar's test is designed for 2 x 2 contingency test (χ^2) analyses of dichotomous variables - in this case, 0 NATA reports versus $\geq$ 1 NATA reports).

TABLE 5 - Predictive statistical and model tuning parameters of the recursive partitioning analyses presented in Figure 1, which interrogated PPLN electrolyte (bias) data (and GGT control) for the best predictors of NATA CMO class **(a)** all electrolyte bias predictors + GGT bias, **(b)** Model with the top three electrolyte - GGT bias predictors from Figure 1 (Calcium, Phosphate, GGT biases).

*Features and Results from RF - PLS Analyses** ***(a)***	**Random Forest (RF)**			
	Optimal mtry	**Accuracy**	**Kappa**	**Final Model (OOB Error Rate)**
Tuned Model (RF)	2	0.648	0.301	**32.14%** <u>No NATA Reports</u> - 100/154 (0.351) <u>NATA CMO Reported</u> - 109/154 (0.292)
*Features and Results from PLS Models**	**Partial Least Squares (PLS)**			
	ROC	**Sensitivity**	**Specificity**	**Final Model Accuracy**
Model Tuning	0.631	0.605	0.583	0.685 (95% CI: 0.580, 0.778)
Final Model Statistics	**McNemar's**	**Sensitivity**	**Specificity**	
	0.026	0.544	0.826	
	Kappa	**PPV**	**NPV**	
	0.370	0.758	0.644	

*Features and Results from RF - PLS Analyses** ***(b)***	**Random Forest (RF)**			
	mtry	**Accuracy**	**Kappa**	**Final Model (OOB Error Rate)**
Tuned Model (RF)	2	0.638	0.2778	**35.06%** <u>No NATA Reports</u> - 97/154 (0.370) <u>NATA CMO Reported</u> - 103/154 (0.331)
*Features and Results from PLS Models**	**Partial Least Squares (PLS)**			
	ROC	**Sensitivity**	**Specificity**	**Final Model Accuracy**
Model Tuning	0.618	0.568	0.628	0.663 (95% CI: 0.557, 0.7583)
Final Model Statistics	**McNemar's**	**Sensitivity**	**Specificity**	
	0.012	0.500	0.826	
	Kappa	**PPV**	**NPV**	
	0.326	0.742	0.623	

R Package - caret algorithms for RF and PLS

All results except mtry, McNemar's ($p < 0.05$ significance) and OOB error rate are presented as a value between 0.0 - 1.0, where a value of 1.0 is perfect accuracy (100%), specificity, or sensitivity etc.

OOB = Out-of-Bag; mtry = number of trees tested at each decision tree in the random forest; ROC = Receiver Operating Characteristic; PPV = Positive Predictive Value; NPV = Negative Predictive Value. NATA CMO - NATA inspection reported Conditions, Minor conditions, Observations. See Methods for RCPAQAP electrolytes bias calculation, details of the R recursive partitioning packages and definitions of the statistical tests included.

The single decision tree (Figure 1c), which was “pruned” by adjusting the model complexity parameter (cp = 0.01, minimum split of 30 per decision node) emphasised the importance of calcium bias as a predictor of NATA class, with a threshold calculated at (+) 0.0017 bias when compared to the RCPAQAP target value. The accuracy of a NATA 0 prediction was 72.3% (81/112), and a correct NATA 1 prediction of 62.8% (123/196), indicating that < 0.0017 calcium bias is more effective for the prediction of NATA 0 cases, in comparison to NATA 1 class prediction above this decision threshold.

In summary, the recursive partitioning modelling via single decision tree and random forest did not produce reliable models, as reflected by contradicting sensitivity/specificity results and the Kappa and McNemar’s statistics. While unreliable models, of the range of electrolytes available as data from the RCPAQAP process, serum calcium and phosphate bias were ranked as the top predictors of NATA class, with control bias marker GGT the third most effective, and more powerful in prediction than 5 other electrolyte RCPAQAP markers combined (removing GGT and/or using lower ranked predictors resulted in a 5 - 10% reduced model predictive accuracy - results not shown).

Based on these results, further machine learning investigations were conducted with the top 2 RCPAQAP electrolyte bias predictor variables, and GGT, using support vector machines (SVM). The SVM is a different algorithm from recursive partitioning. Rather than calculating model entropy or a Gini coefficient, the SVM calculates a separating hyperplane from “support vectors”, weighted within the model to separate complex data in a higher dimension of computational space (6). Whether SVM produces more accurate and robust models from the same results (Figure 1, Table 5) is examined in the following section.

Support Vector Machines - The complexity of the relationship between RCPAQAP marker bias and NATA categories decided by the number of C, M and/or O reports, is amply demonstrated in Figure 2. The SVM allows a “slice” through a higher dimension in relation to x and y variables, characterising the relationship with the NATA class (0 or 1) response variable. For Figure 2, all plots had calcium bias and phosphate bias on the x - y axes, with the slice representing GGT at different bias levels (0.0 to - 0.20), revealing different patterns in the NATA class 1 results (positive GGT slices did not produce a dichotomy between NATA classes 0 and 1 - only negative GGT bias generated plots representing these classes).

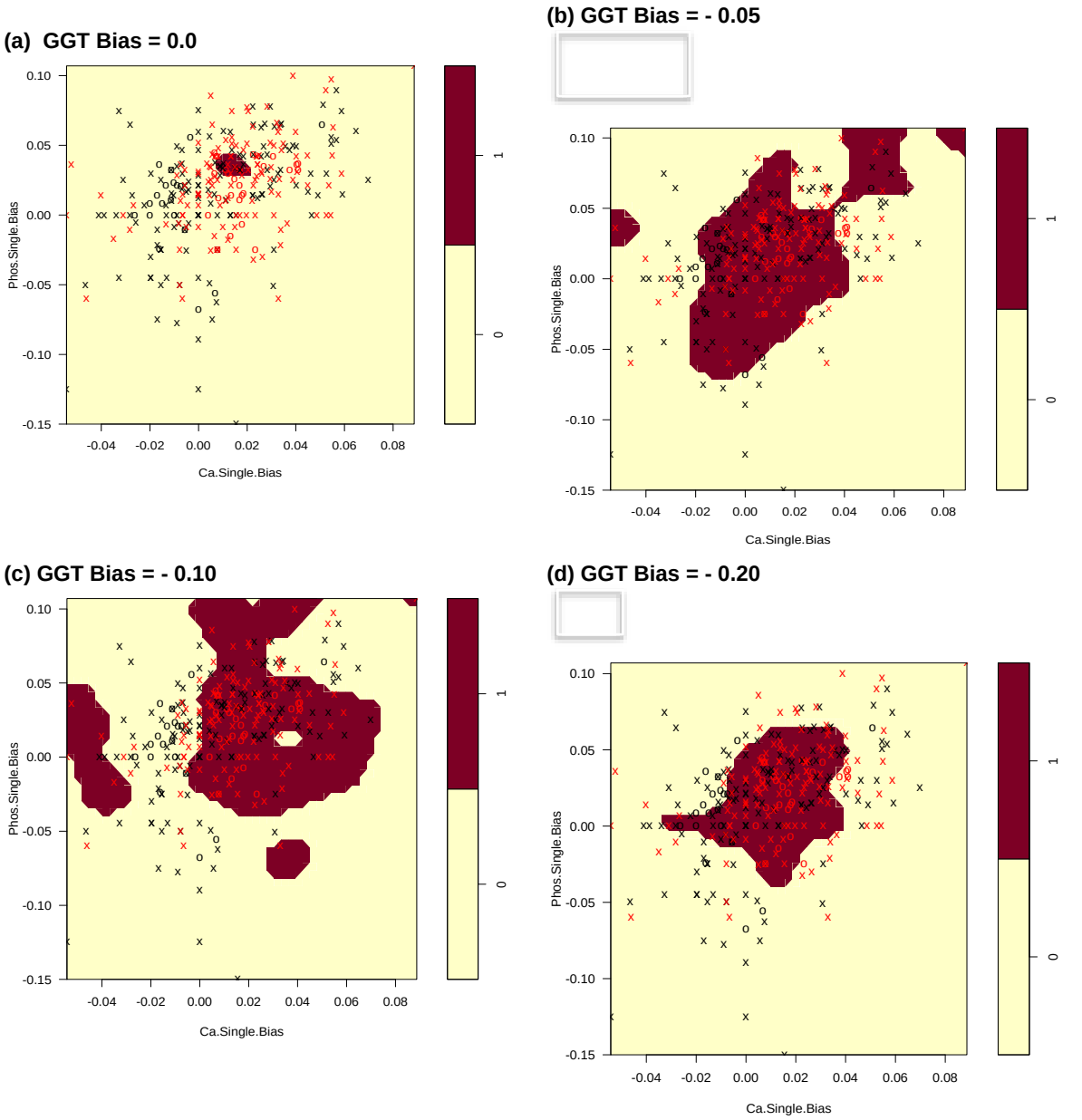


Figure 2 - Support Vector Machine (SVM) plots summarising the relationship between serum phosphate and calcium (total) RCPAQAP bias, in the context of a constant GGT RCPAQAP bias, to predict PPLN laboratories with $\geq$ 1 NATA Condition, Minor report, or Observation (NATA Class 1). GGT bias held constant at **(a)** 0.0; **(b)** - 0.05; **(c)** - 0.10; **(d)** - 0.20. Rectangle grid in **(b)** and **(d)** inserted to estimate example prediction rules for PPLN laboratories with NATA reports.

Results: At a GGT bias value of 0.0 (representing where GGT target values were perfectly attained by the laboratory during the RCPAQAP cycle), class 1 NATA laboratories were barely detectable on the SVM plot, with a narrow range of 0.03 - 0.05 phosphate and 0.0 - 0.02 calcium RCPAQAP bias (Figure 2a). In the range of GGT (-) 0.05 to (-) 0.20 bias (laboratory GGT results under the RCPAQAP target value), class 1 NATA laboratories are clearly detectable in the SVM plots (Figs 2b - d).

A feature of the Phosphate - Calcium - GGT (RCPAQAP bias) SVM model was the fragmentation of NATA class 1 clusters into separate islands, emphasising the non-uniform nature of the relationship between RCPAQAP and NATA results (Figs. 2b - c). In terms of

size, the SVM plot with the GGT bias slice of - 0.20 was smaller (Fig. 2d), suggesting that decreasing the GGT slice further would lead to non-detectable class 1 clusters; thus, the effective range for determining the phosphate + calcium RCPAQAP biases for PPLN laboratories with NATA reports recorded was GGT (RCPAQAP) bias levels between - 0.05 to an approximate lower limit of - 0.30.

TABLE 6 - Summary of Figure 2 SVM plots to predict NATA Class by Ca.Bias, Phos.Bias, GGT.Bias

SVM Model	Method and Kernel	Tuning & Statistical Coefficients		Accuracy (%) (Range)
		Gamma	Cost	
Full Model *	C-classification Kernel = Radial	2	1	66.56 (51.61 - 83.87)
Train & Test #	C-classification Kernel = Radial	Gamma	Cost	Accuracy (Diag. %) - Correct Class Prediction
		2	1	59.82 Class (N) - 28/38 (73.7%) Class (Y) - 27/54 (50%)
		Kappa	Rand	
		0.22	0.51	

* 10-fold cross-validation on training data

R Package (caret) tuning and testing: 70 - 30% training/testing data split. Accuracy (Diagonal %): Calculated from the major diagonal of the 2 x 2 contingency table of correct or incorrect predictions ("confusion matrix"). Rand (Index) (- 1.0 to 1.0): How well the trained SVM model predicts *True Positives*, *True Negatives*, *False Positives*, *False Negatives*.

Class (N) = No NATA Reports; Class (Y) = NATA Reports recorded ($\geq$ 1 per lab in the sample).

From the SVMs, examples of NATA Class 1 prediction models are (Fig. 2b+d: *Rectangle*) - (1) **NATA Class 1** = GGT (- 0.05) + Phosphate (- 0.01 ↔ 0.03) + Calcium (- 0.018 ↔ 0.04) (2) **NATA Class 1** = GGT (- 0.20) + Phosphate (0.04 ↔ 0.06) + Calcium (0.01 ↔ 0.035).

To develop precise rules like (1) and (2) from SVM models can continue by inserting grids of various shapes into the plots and extrapolating to the x - y axes to estimate the phosphate and calcium bias ranges, captured at a specific GGT bias value. One can also generalise the plots; for example, Figure 2c (GGT Bias - 0.10) shows three distinct areas representing NATA Class 1 laboratories, with a cluster at (0.0 ↔ 0.05) phosphate bias + (- 0.03 ↔ - 0.06) calcium bias, as well as larger and smaller clusters of similar phosphate bias, but calcium biases from 0.0 ~ 0.08. To use these rules efficiently, each laboratory must conduct further testing and validation of RCPAQAP samples to optimise performance and feedback specific to their laboratory environment, particularly in terms of the analytical platforms.

However, as found for the recursive partitioning models, the measures of SVM model robustness also indicated poor predictive performance (Table 6). The kappa statistic of 0.22 suggests that prediction agreement occurred 22% of the time, and Rand Index (that likewise

measures agreement from 2 x 2 confusion matrix) says that the correct prediction of true positives (TP), false negatives (FN) etc was only successful in 51% (0.51) of cases (Table 6). Therefore, caution must be applied when interpreting these results, and as suggested above, require further optimisation through the input of more RCPAQAP data, and repeated testing and validation. The results presented herein, provide a guide only since the rules were calculated from one RCPAQAP - NATA cycle.

PPLN Electrolytes - Conclusion: The simplest rule to interpret for the prediction of PPLN laboratory NATA Class was the single decision tree (Figure 1c). The optimised decision tree ("pruned") is an example of the application of a recursive partitioning algorithm, and provides a rule solely on calcium RCPAQAP bias, with a threshold of > or < 0.0017 as the decision point to separate NATA classes (or 0.0017 can be assigned a value of 0.0 for simplicity).

Moving to a different, more powerful algorithm, the SVM, demonstrated greater range and complexity for the decision rules, although these did accord broadly with the decision tree calcium bias threshold by including 0.0. The statistics applied to both recursive partitioning (tree) and SVM models indicated a lack of robustness, but provides a guide for further optimisation with the addition of data from future RCPAQAP - NATA quality cycles.

(ii) PPLN Liver Function Tests (LFTs) - The same analytical logic was applied to the interrogation of LFT marker bias versus the same NATA reports for PPLN laboratories, as presented above for PPLN electrolyte RCPAQAP via recursive partitioning and SVMs.

NATA Classes representing all CMO reports, or M reports only, feature in LFT RCPAQAP bias investigations. This was required since the quality of NATA Class prediction results varied depending upon the number of RCPAQAP bias variables included; for example, the prediction of total CMO Class was best with only the top 5 LFT.Bias RF variables (otherwise, NATA Class 1 predictions had an error > 50%). NATA M Class only models produced improved models with all LFT.Bias markers included, although this changed again depending on how RF models were constructed. The results presented hereafter were deemed the best representatives from the total investigation, based, for example, on model metrics like McNemar's statistic.

Recursive Partitioning - Figure 3 summarises RF and decision tree (recursive partitioning) models for NATA CMO (total) class predictions (a and b), and a decision tree model for the prediction NATA M Classes (c).

The top five LFT (bias) predictors separating PPLN laboratories that recorded no NATA CMO reports from laboratories with 1 or more NATA reports were (in descending order of importance); LD, AST, ALP, Albumin and GGT.

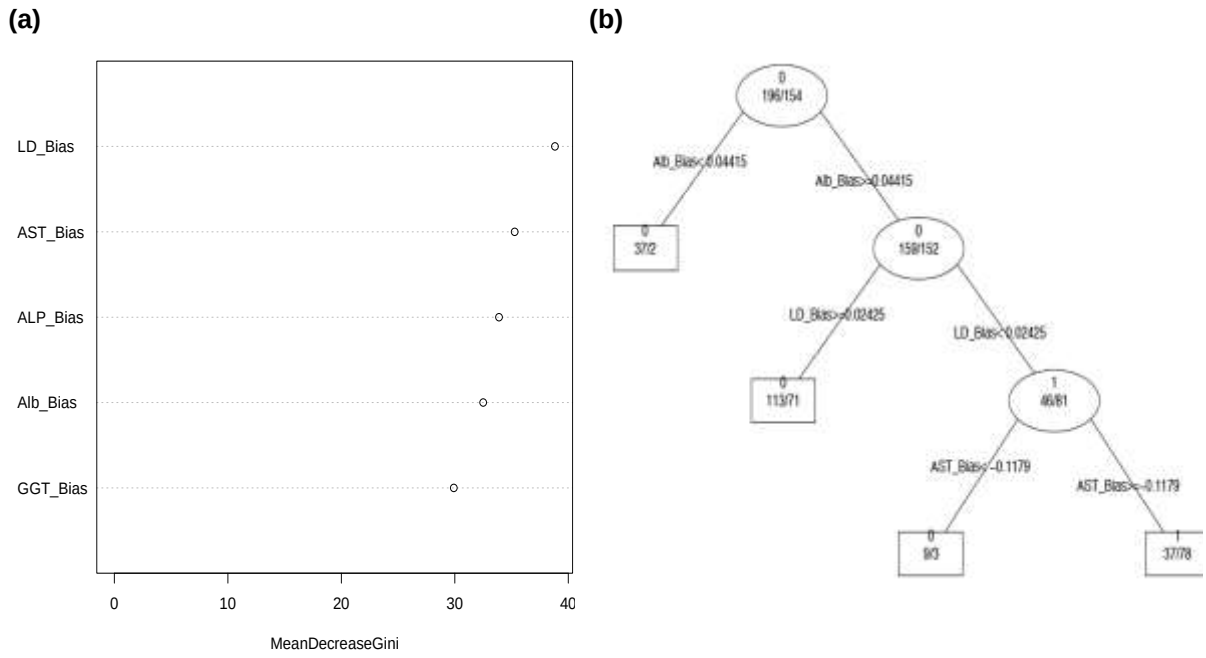


Figure 3 - PPLN RCPAQAP LFT bias predictions of NATA class by recursive partitioning - **(a)** Full Random Forest Analysis (RFA) of the top 5 RCPAQAP LFT bias variables in relation to NATA CMO Classes; **(b)** Single Decision Tree on the identical variables used for the RFA model presented in (a) (Pruned = minsplit = 30, cp = 0.030); **(c)** Single Decision tree constructed with all available LFT RCPAQAP bias markers, in relation to NATA M Class alone. (Pruned minsplit = 30, cp = 0.030).
** NATA Classes decided as the laboratories that recorded NATA C, M and/or O reports (1), compared to laboratories with no NATA reports recorded (0), which coincided with the RCPAQAP cycle investigated (C - major condition; M - minor condition; O - Observation).*

For NATA CMO Class prediction, the inclusion of ALT, TBil, DBil and TP biases in some models resulted in Class 1 prediction error rates of greater than 50%, hence the focus on a subset of LFT bias predictors available for modelling.

TABLE 7 - Summary of the PPLN-LFT decision tree model presented in Figure 3c (NATA M Class prediction: n= 355 - Root and terminal nodes only). Best Minor 0 or 1 predictions are highlighted.

Split	n	loss	y-value	y-probability *	
				Correct	Incorrect
Root	355	138	0	0.611	0.389
LD.Bias ≥ 0.029	182	48	0	0.736	0.264
Alb.Bias < 0.044	20	2	0	0.900	0.100
GGT.Bias < - 0.130	13	2	0	0.846	0.154
ALP.Bias ≥ 0.036	12	3	0	0.750	0.250
ALP.Bias < 0.036	93	39	1	0.581	0.419
AST.Bias ≥ - 0.061	35	6	1	0.829	0.171

* The y-probability indicates the accuracy of predicting either NATA Minor Class 0 (No Minor reports) or NATA Minor Class 1 (≥ 1 Minor reports) (denoted above by the y-value column).

Considering the prediction LFT rules presented in Fig. 3c, and results in Table 7, the following decision rules are proposed to predict NATA outcome for PPLN laboratories:

Decision tree rules (Table 7 and Figure 3c) -

(3) LD.Bias (< 0.029) + Albumin.Bias (< 0.044) = NATA M Class 0 (90% correct)

(4) LD.Bias (< 0.029) + Albumin.Bias (≥ 0.044) + AST.Bias (≥ - 0.061) = NATA M Class 1 (83% correct)

The recursive partitioning models summarised by Figure 3a-b (total CMO NATA Classes) were similar to Figure 3c (M only NATA Classes), except that the role of ALB.Bias (serum albumin) as a leading predictor varied. The random forest (RF) identified ALP.Bias as a more important variable than ALB.Bias, with AST.Bias and LD.Bias important for all models.

Identical to the PPLN electrolyte investigations, the RF models were evaluated for accuracy and robustness. Table 8 summarises models used to predict NATA classes based on the presence or absence of total NATA reports (C, M and O) (Table 8b), and the presence or absence of NATA M only reports (Table 8a).

The best model was the RF that included all LFT.Bias variables to predict NATA M Classes (Table 8a). A model accuracy of 70.2% was achieved, with a McNemar's statistic > 0.05, indicating an effective prediction of TP, TN, FP, FN from the prediction ("confusion") matrix, with a Kappa result for the model of 0.351. Therefore, while not ideal, a prediction model based on the rules from the corresponding decision tree (Figure 3c) was confirmed (Decision rules 3 & 4). The model analyses from Table 8a also display reasonable positive and negative predictive values for NATA Minor Class model, also as found previously for

PPLN electrolytes, specificity was more effective, although not significantly for this specific RF model.

The superiority of the RF (train - test) model summarised in Table 8a was reinforced by the less impressive results from the PLS model on the same integrated NATA - RCPAQAP results (all RCPAQAP LFT bias predictors versus NATA M Class - also Table 8a). McNemar's statistic was significant ($p < 0.05$) indicating that model predictions were significantly altered in comparison to the pre-prediction classes, which again on inspection of the results was due to the poor performance in predicting true positives (TP). Kappa was also less (< 30% agreement), with final predictive values 3 - 5% lower for the PLS modelling.

RF and PLS models summarised in Table 8b were focussed on the prediction of NATA Classes derived from the total C, M reports and Observations by the top five RCPAQAP bias predictors (Figure 3a). This time the PLS model was slightly superior to the RF model as judged by the ROC results, but in general both approaches to modelling NATA and integrated RCPAQAP bias results were poor, as assessed via McNemar's and Kappa statistics.

TABLE 8 - Predictive statistical and model tuning parameters of the recursive partitioning analyses presented in Figure 3, which interrogated PPLN LFT (bias) data for the best predictors of NATA class - as assessed via caret Random Forest or Partial Least Squares. **(a)** All LFT bias predictors to separate NATA classes were based solely on M (minor) NATA conditions: **(b)** Model with the top five LFT bias predictors from Figure 3a (total CMO NATA reports). *Continued on Page 29*.

*Features and Results from RF Models** ***(a)***	**Random Forest (RF)**			
	Optimal mtry	**Sensitivity**	**Specificity**	**Final Model Accuracy**
Model Tuning	3	0.427	0.763	0.702 (95% CI: 0.604, 0.788)
Final Model Statistics	**McNemar's**	**Sensitivity**	**Specificity**	
	0.151	0.512	0.825	
	Kappa	**PPV**	**NPV**	
	0.351	0.656	0.722	
*Features and Results from PLS Models**	**Partial Least Squares (PLS)**			
	ROC	**Sensitivity**	**Specificity**	**Final Model Accuracy**
Model Tuning	0.709	0.433	0.805	0.673 (95% CI:0.574, 0.762)
Final Model Statistics	**McNemar's**	**Sensitivity**	**Specificity**	
	0.026	0.415	0.841	
	Kappa	**PPV**	**NPV**	
	0.272	0.629	0.688	

→

*Features and Results from RF-1 Models** ***(b)***	**Random Forest (RF) - 1**			
	Optimal mtry	**Accuracy**	**Kappa**	**Final Model (OOB Error Rate) - Figure 3a**
Tuned Model (RF)	5	0.596	0.174	**35.55%** No NATA Reports - 129/192 = 0.33 NATA CMO Reported - 94/154 = 0.39
*Features and Results from RF-2 Models**	**Random Forest (RF) - 2 (mtry = 2)**			
	ROC	**Sensitivity**	**Specificity**	**Final Model Accuracy**
Model Tuning	0.639	0.512	0.683	0.615 (95% CI:0.515, 0.709)
Final Model Statistics	**McNemar's**	**Sensitivity**	**Specificity**	
	0.040	0.413	0.776	
	Kappa	**PPV**	**NPV**	
	0.195	0.594	0.625	
*Features and Results from PLS Models**	**Partial Least Squares (PLS)**			
	ROC	**Sensitivity**	**Specificity**	**Final Model Accuracy**
Model Tuning	0.724	0.531	0.751	0.625 (95% CI: 0.525, 0.718)
Final Model Statistics	**McNemar's**	**Sensitivity**	**Specificity**	
	0.025	0.413	0.793	
	Kappa	**PPV**	**NPV**	
	0.213	0.613	0.630	

Support Vector Machines - The best recursive partitioning model was explored further by SVM interrogation. In short, the tuned SVM models confirmed the results from the recursive partitioning analyses, with overall predictive accuracies at approximately 68%, and greater success at predicting the NATA Low (zero reports) Class than the Class representing NATA conditions or observations (69.6% versus 62.5%; Table 9).

Figure 4 shows representative SVM plots for LD.Bias versus AST.Bias with an ALP.Bias slice at - 0.20 (a), 0.0 (b) and 0.02 (c) bias values. ALP.Bias was identified by recursive partitioning (Figure 3) as of similar predictive power as ALB.Bias, which was featured in the decision support model proposed earlier (ALB.Bias was chosen for decision support since it provided terminal decision nodes with more cases available for an accuracy calculation - Figure 3a - Table 7).

In addition, the ultimate predictive values of the RF and PLS models were 5 - 10% less than the Table 8a (RF) model. Table 8b also presents a standard tuned RF model (RF - 1) that according to the OOB error rate, had a < 65% accuracy, and a Kappa value < 20%.

On this basis, the final model of NATA Class prediction will use Minor reports only from a model that included all RCPAQAP (LFT) bias variables available, with the final results summarised by Figure 3c and decision rules (3) and (4).

TABLE 9 - Summary of Figure 4 SVM plots, reporting the results of the full PPLN LFT (RCPAQAP Bias) SVM Model for Minor NATA responses.

SVM Model	Method and Kernel	Tuning & Statistical Coefficients		Accuracy (%) (Range %)
		Gamma	Cost	
Full Model *	C-classification Kernel = Radial	0.0313	16	68.391 (57.14 - 80.0)
Train & Test #	C-classification Kernel = Radial	Gamma	Cost	Accuracy (Diag. %) - Correct Class Prediction
		0.0313	16	68.103 Class (N) - 64/92 (69.6%) Class (Y) - 15/24 (62.5%)
		Kappa	Rand	
		0.2481	0.562	

* 10-fold cross-validation on training data

R Package (caret) tuning and testing: 70 - 30% training/testing data split. Accuracy (Diagonal %): Calculated from the major diagonal of the 2 x 2 contingency table of correct or incorrect predictions ("confusion matrix"). Rand (Index) (- 1.0 to 1.0): How well the trained SVM model predicts *True Positives*, *True Negatives*, *False Positives*, *False Negatives*.

Full model: NATA Minor Class ~ Alb.Bias + ALP.Bias + AST.Bias + GGT.Bias + LD.Bias. Final SVM prediction models used LD.Bias + AST.Bias + ALP.Bias (ALP = slice of fixed bias value) only to predict NATA M class. Class (N) = No NATA M Reports; Class (Y) = NATA M Reports recorded ($\geq$ 1 per lab in the sample).

The SVM patterns for PPLN-LFTs had clearer trends in comparison to PPLN-electrolyte SVM plots (Figure 2). For the ALP.Bias slice at - 0.20 (Fig. 4a) the entire range of LD.Bias (- 0.06 to 0.10) was involved in the prediction of PPLN laboratories that received NATA conditions or observation reports (NATA Class Y - denoted as "High" in the plots), with AST.Bias reducing on a steady gradient from a peak of 0.0 (LD.Bias: - 0.06), to approximately an AST.Bias of - 0.15 at a corresponding LD.Bias value of 0.04 - 0.05, after which the AST.Bias increased to approximately - 0.12 (LD.Bias 0.10). At ALP.Bias slices of 0.0 and 0.02 (Figures 4b - c), similar NATA Class Y predictions were observed with ranges between approximately - 0.03 to - 0.20 for AST.Bias, and - 0.05 to - 0.01 for LD.Bias. Extending the ALP.Bias slice to smaller or larger bias values did not produce a clear separation between NATA Y and N Classes (results not shown). Furthermore, the SVM results suggested that for laboratories that attracted NATA Minor conditions, prediction results under the designated RCPAQAP target value were achieved as a general trend.

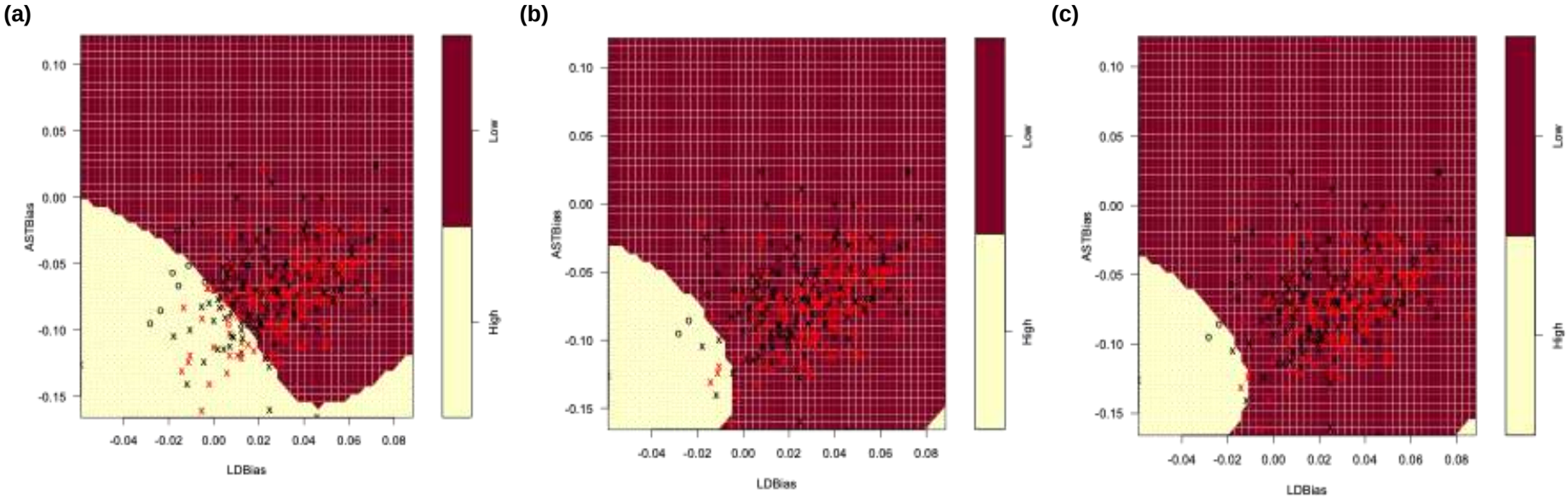


Figure 4 - Support Vector Machine plots representing the patterns associated with PPLN laboratories that received NATA reports and observations (Class - High: Yellow) or reported zero conditions and observations (Class - Low: Maroon), with model features summarised in Table 9.

The NATA Class prediction model was:

NATA Minor (M) Class (High or Low) ~ Alb.Bias + ALP.Bias + AST.Bias + GGT.Bias + LD.Bias. Plots present the interaction of LD.Bias + AST.Bias + ALP.Bias to predict NATA M class. RCPAQAP AST.Bias sits on the y-axes, and LD.Bias on the x-axes, with ALP.Bias applied to slice the AST - LD plots at RCPAQAP bias values of **(a)** - 0.20; **(b)** 0.0; and **(c)** 0.02 ALP.Bias.

Conclusions (PPLN - LFT): Recursive partitioning produced a robust NATA Class prediction of 70% accuracy, with positive and negative predictive values of 66% and 72% respectively, generated from a model comprising only Minor (M condition) NATA report response classes, and all LFT RCPAQAP bias variables. While relatively robust as indicated by the McNemar's statistic, the prediction of true negatives (TN) was favoured, which impacted model performance as shown by the Kappa results (Table 8a). RF algorithms performed better in comparison to PLS (R caret package training and testing), with decision trees providing support by calculating LFT.Bias thresholds (Figure 3c).

SVM investigations of the same PPLN - LFT bias data did not produce robust models (e.g. Rand Index), with predictive accuracies of less than 70%. From these machine learning models, the RCPAQAP LFT biases of LD, AST, ALP and albumin (ALB) were best for the prediction of which PPLN laboratories received NATA condition reports and/or observations. SVM results suggested that in general PPLN laboratories with NATA reports came under the RCPAQAP target values, with more data required to investigate greater model sensitivity.

(iii) SPLN Electrolytes - Recursive partitioning and SVM investigations were conducted for integrated NATA and RCPAQAP results, exactly as performed for PPLN laboratories.

As explained in section 3 (a), the profiles of NATA results were very different for PPLN in comparison to SPLN. For the PPLN laboratory NATA profile, conditions, minor reports and observations were counted to provide enough scope to create two NATA response classes for interrogation. The result was a Class with no NATA reports, and the "yes" class consisting of laboratories with 1 or more conditions/minor reports or observations (C, M, O). For SPLN, 128 Conditions (C) and 244 Minor total NATA reports were recorded (Table 2). With this NATA C - M profile, the count of observations was not required. The NATA response classes for interrogation by SPLN electrolyte or LFT RCPAQAP bias, therefore, were calculated as above or below the median number of NATA C-M reports (the "low" class contained laboratories with zero reports, but these were a minority). As warned earlier in the report, this situation made a direct comparison of eventual PPLN and SPLN machine learning models difficult, but nonetheless will assist in identifying features of laboratories with varied NATA performance, as captured by the RCPAQAP cycle.

The SPLN NATA - RCPAQAP recursive partitioning and SVM results follow, for electrolytes (+ GGT) and LFT results, again represented as aggregated RCPAQAP relative bias results to predict NATA response (High versus Low) Classes.

Recursive Partitioning - The random forest (RF) and decision tree investigations successfully validated two previous results observed for SPLN laboratories: (1) a focus on the NATA reporting of minor (M) conditions only, and not a combination of major and minor conditions (C & M) as a response for RCPAQAP bias modelling - these produced the most accurate

and robust integrated predictive models, and (2) GGT.Bias again showed a powerful interaction with RCPAQAP electrolyte bias (2, 3) in the prediction of (M) NATA Classes.

TABLE 10 - Predictive statistical and model tuning parameters of the recursive partitioning analyses presented in Figure 5, which interrogated SPLN electrolyte + GGT (bias) data for the best predictors of NATA class (High or Low NATA Minor condition reports), as assessed via caret Random Forest or Partial Least Squares.

*Features and Results from RF Models**	**Random Forest (RF)**			
	Optimal mtry	**Sensitivity**	**Specificity**	**Final Model Accuracy**
Model Tuning	4	0.735	0.647	0.712 (95% CI: 0.641, 0.776)
Final Model Statistics	**McNemar's**	**Sensitivity**	**Specificity**	
	0.272	0.771	0.648	
	Kappa	**PPV**	**NPV**	
	0.420	0.705	0.722	
*Features and Results from PLS Models**	**Partial Least Squares (PLS)**			
	ROC	**Sensitivity**	**Specificity**	**Final Model Accuracy**
Model Tuning	0.715	0.719	0.626	0.674 (95% CI: 0.601, 0.741)
Final Model Statistics	**McNemar's**	**Sensitivity**	**Specificity**	
	0.028	0.781	0.557	
	Kappa	**PPV**	**NPV**	
	0.341	0.658	0.700	

Model interrogated - NATA Minor Class → RCPAQAP Bias results: serum Bicarbonate + Calcium + Chloride + Creatinine + Magnesium + Phosphate + Potassium + Sodium + (GGT) Bias.

Deeper analysis of the recursive partitioning models for integrated NATA - RCPAQAP results are summarised in Table 10, with supporting plots displayed in Figure 5. The model evaluated the prediction of NATA M Classes by bias values calculated from all electrolyte markers, and GGT, run for RCPAQAP assessment.

The tuned RF (caret) model had an overall accuracy prediction of 71.2%, a non-significant McNemar's statistic (p = 0.272), with both positive and negative predictive values over 70% (Table 10). Sensitivity was approximately 9 - 14% higher than specificity, indicating greater success at correctly predicting true positive results. The Kappa statistic was less than 50%, and while an improvement on the same statistic for PPLN laboratories, suggests poor performance in prediction agreement with the algorithm. Table 10 also presents the caret PLS results of the same RCPAQAP + NATA model. Poorer performance was found in comparison with the RF model, with a significant McNemar's results and lower Kappa. Overall accuracy was less, as well as for predictive values, particularly the positive predictive value (PPV). The role of GGT.Bias was also evaluated for the identical recursive partitioning models (Figures 5a,5b).

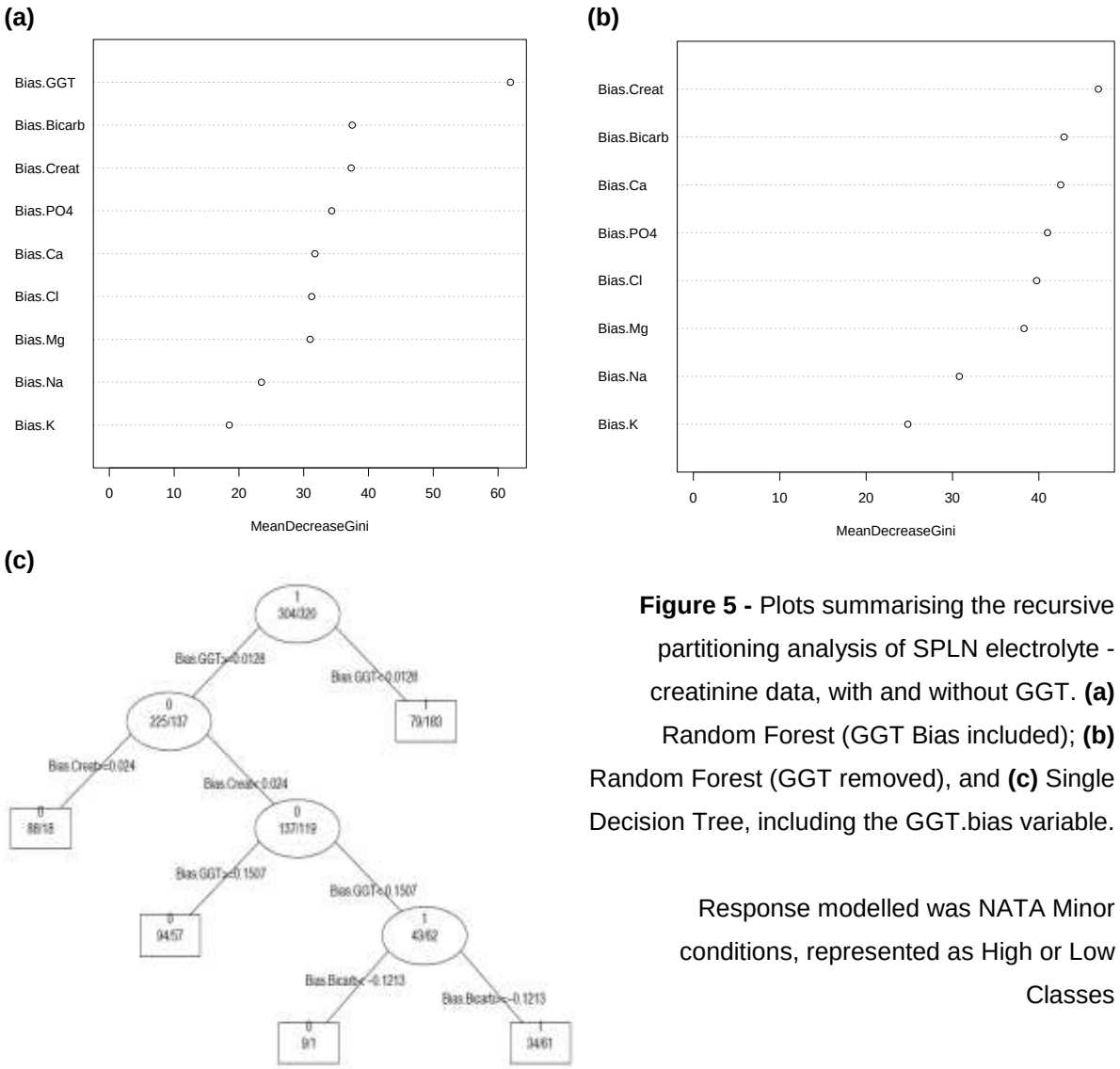


Figure 5 - Plots summarising the recursive partitioning analysis of SPLN electrolyte - creatinine data, with and without GGT. **(a)** Random Forest (GGT Bias included); **(b)** Random Forest (GGT removed), and **(c)** Single Decision Tree, including the GGT.bias variable.

Response modelled was NATA Minor conditions, represented as High or Low Classes

The removal of GGT.Bias weakened the model accuracy, particularly the prediction of the low (M) NATA Class, although both NATA Classes were less accurate (Table 11).

ANCOVA modelling of GGT.Bias (dependent variable) in relation to fixed factor(s) (e.g. NATA Class) and continuous covariates showed that the top-ranked UEC.Bias markers significantly explained the variation of GGT.Bias, for both the NATA Minor Class and the NATA combined Minor & Condition Class (see Performance Report 4 - Table 2). The leading RF predictors (e.g. bicarbonate bias, creatinine bias) had highly significant p - values (i.e. $p < 10^{-9}$), while potassium and sodium bias were not significant covariates to explain GGT.Bias. NATA Minor alone and Minor + Condition Class models both had R^2 values of 0.76, indicating that the variables in the model explained 76% of GGT.Bias variation. Of additional interest was that dividing the SPLN laboratories into B or G categories was not significant in terms of this RCPAQAP and NATA integration model, indicating no effect for laboratory category.

TABLE 11 - The impact of GGT.Bias calculated from RCPAQAP data on the Electrolytes, Creatinine (EC) model of NATA (Minor report) Class random forest modelling and prediction (Fig. 5).

NATA Class - Minor Reports	GGT Included in UEC Model		GGT Excluded from UEC Model	
	Prediction Accuracy (%)	Overall Accuracy (%)	Prediction Accuracy (%)	Overall Accuracy (%)
High (> Median)	241/320 (75.3%)	71.3%	230/320 (71.9%)	66.4%
Low (< Median)	198/296 (66.9%)		179/296 (60.5%)	

Model interrogated - NATA Minor Class → RCPAQAP Bias results: serum Bicarbonate + Calcium + Chloride + Creatinine + Magnesium + Phosphate + Potassium + Sodium + (GGT) Bias.

The median number of Minor (condition) reports across all SPLN laboratories was calculated to allow the designation of high (> median) and low (< median) NATA Classes for modelling with RCPAQAP results.

GGT Bias was also the leading predictor when included in the LFT.Bias panel to explain NATA classes for SPLN laboratories (next section - iv), while it was not a leading predictor for PPLN results.

Therefore, a key finding from this research programme (this and the previous pilot study) was the identification of GGT as the leading RCPAQAP marker with which to understand patterns in NATA reports, but it seems, only for laboratory networks that attract a broad range and diversity of NATA Conditions and Minor (conditions) reporting. As shown by the electrolyte (UEC) results, this predictive power is useful broadly, not only for LFT profiles (an explanation for the GGT impact is presented in section (4), below).

An analysis identical to that presented in Table 7 was conducted, from which a predictive model can be proposed. The models were designed to take the best decision tree terminal node predictions (Fig. 5) with the least loss of cases. Therefore, to predict the High (Minor) NATA Class, we have the following decision rules (Table 10, Figure 5);

(5) GGT.Bias (≥ 0.0128) + Creatinine.Bias (≥ 0.024) = NATA Low Class (83.02%)
(6) GGT.Bias (< 0.0128) = NATA High Class (69.85%) *

* The next most accurate decision rule for NATA High (64.21%) involved creatinine and bicarbonate biases, as well as GGT.

Support Vector Machines - As conducted for the PPLN NATA - RCPAQAP analyses, SVM was also applied to the results to assess whether a model of increased accuracy/utility could be identified via this algorithm.

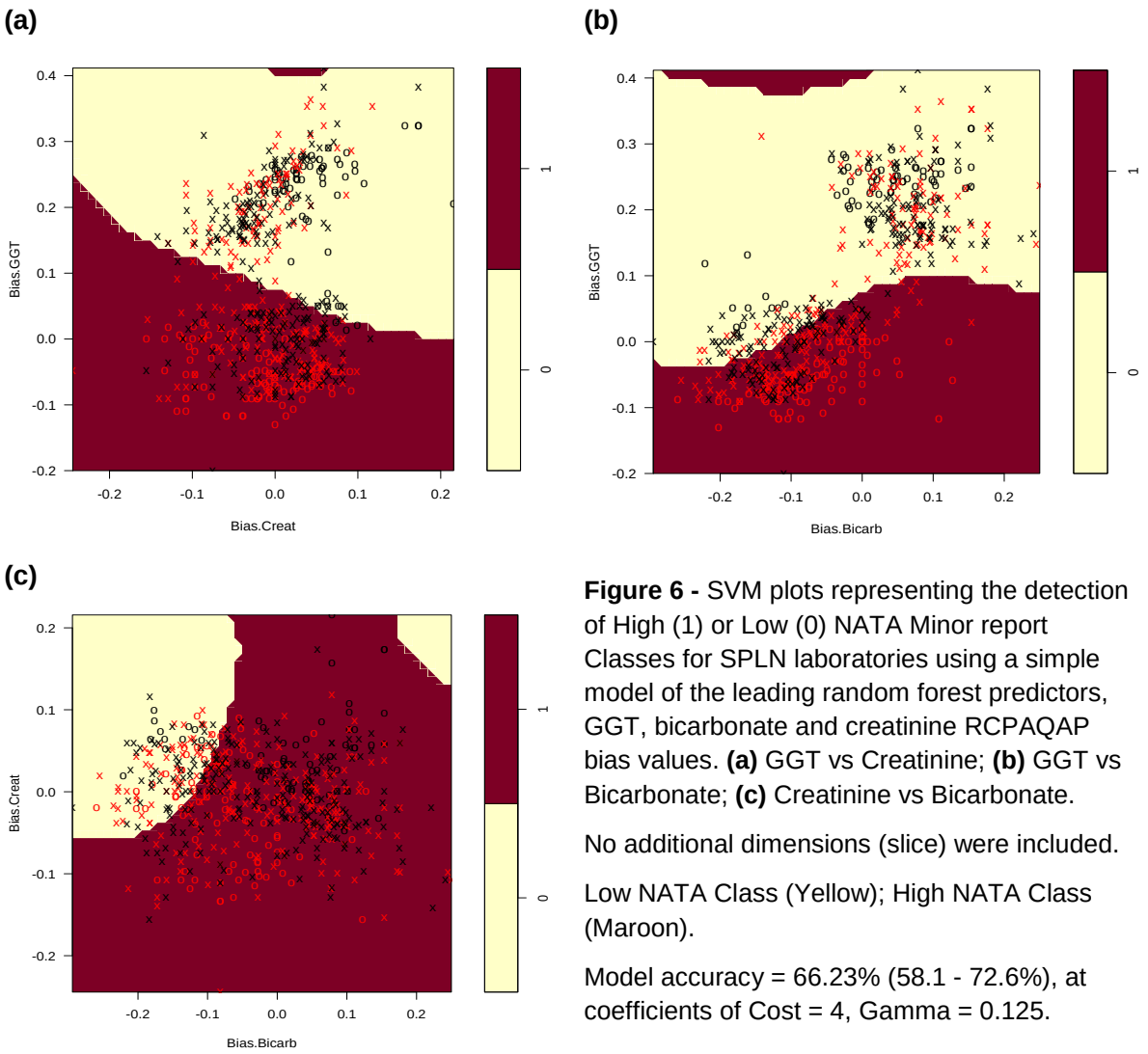


Figure 6 - SVM plots representing the detection of High (1) or Low (0) NATA Minor report Classes for SPLN laboratories using a simple model of the leading random forest predictors, GGT, bicarbonate and creatinine RCPAQAP bias values. **(a)** GGT vs Creatinine; **(b)** GGT vs Bicarbonate; **(c)** Creatinine vs Bicarbonate.

No additional dimensions (slice) were included.

Low NATA Class (Yellow); High NATA Class (Maroon).

Model accuracy = 66.23% (58.1 - 72.6%), at coefficients of Cost = 4, Gamma = 0.125.

Figure 6 summarises the interaction of the top three NATA Class predictors as determined by RF (Figure 5a). No slice dimensions were applied to these models due to the difficulties of defining plots with separated high versus low NATA Classes. The RCPAQAP bias GGT interactions with creatinine or bicarbonate showed a reversed pattern of GGT fluctuation over a broad range for both RCPAQAP markers (Figures 6 a, b). At approximately - 0.3 creatinine, GGT bias was 0.25, with this trend descending to 0.0 GGT by just over 0.2 creatinine bias. The opposite was true for bicarbonate in relation to GGT, with GGT below 0.0 for an approximate value of - 0.3 bicarbonate, increasing to 0.1 GGT at a bicarbonate value of 0.1 bias, then decreasing slightly. Both bicarbonate and creatinine (Figs. 6a, b) also showed a NATA class region at 0.4 GGT, of different widths, suggesting an additional decision rule.

Of note was the wider range of bias variation among GGT, bicarbonate and creatinine, when compared to the SVM plots presented earlier in the report for PPLN laboratories. As reported previously (Performance/Milestone Report 4 and others), there is an obvious separation of GGT bias for SPLN laboratories associated with NATA Minor Class. The effect was not as subtle as found for other interactions, which therefore can be proposed as a reason for these wider RCPAQAP bias ranges.

The SVM plot describing the interactions of bicarbonate and creatinine bias in predicting NATA Class had two distinct NATA low regions, the first extending from - 0.05 creatinine to approximately - 0.1 bicarbonate bias (Figure 6 c). The second smaller region covered creatinine ~ 0.15 - 0.20 and 0.16 to > 0.20 bicarbonate bias.

Referring to the corresponding decision tree (Figure 5c) showed that the bias ranges detected by SVM were broadly accurate, particularly that involving the GGT.Bias decision tree node of > < 0.1507. The SVM model (Figure 6) had an accuracy of 66.23% (Cost = 4; Gamma = 0.125). No SVM prediction models exceeded an accuracy of 70%, in spite of algorithm ensemble tuning and evaluation (see Appendix E for machine learning ensemble analyses, and SVM model development examples).

Conclusions (SPLN - UEC + GGT) - As alluded to earlier, differences due to the range and variety of NATA reports for PPLN in comparison to SPLN were likely to impact the prediction models and decision rules (differences in sample size between PPLN and SPLN may also contribute - not explored directly for this project).

The SPLN results showed again that in combination with RCPAQAP bias results for UEC (although serum urea was not available), GGT was a powerful predictor of category/class calculated from the count of minor conditions by NATA, with ANCOVA analysis previously showing the strength of the top RF UEC predictors as covariates that explain the variation in GGT bias, calculated from RCPAQAP results. This GGT function in predicting NATA results has now been validated by a second study, and alone contributes an additional 5% accuracy to the RF models (Table 11). For the equivalent UEC study of PPLN laboratories, GGT.bias ranked highly in RF models, but was not as powerful in the PPLN context (although GGT provided a slice for the PPLN SVM models).

RF models were more effective for SPLN analysis in terms of accuracy and model robustness statistics, compared to PLS tuned models, with a trend to being more successful for correctly predicting true positives, which was different trend in comparison to PPLN UECs.

The SVM analyses provide useful plots to explain the complexity of the interactions between 2 - 3 variables as predictors of response class. In the specific case of SPLN, the SVM patterns demonstrated the wide range of creatinine and bicarbonate bias, with GGT.Bias promoting variation in the prediction thresholds (Figures 6 a, b), again demonstrating its powerful influence on NATA Class prediction.

(iv) SPLN LFTs - The SPLN RCPAQAP results for LFTs offered extra markers (e.g. direct bilirubin), but only markers in common with PPLN results were used for the recursive partitioning and SVM models that follow.

GGT.Bias has been established as a powerful individual predictor within an electrolyte profile, so interest in GGT performance with other LFTs will be viewed closely. The available markers cover routine and true LFTs (e.g. ALT and total bilirubin, respectively). This variety of LFT QAP markers will provide opportunities to better understand RCPAQAP interactions with NATA results.

Recursive Partitioning - GGT.Bias was the leading predictor among the RCPAQAP results for the selected LFTs applied to recursive partitioning models (Figure 7a). Also notable were the single decision trees, one designed to detect NATA classes calculated from the number of Minor conditions (Figure 7b), and the other to detect NATA classes derived from Conditions (Figure 7c). The RF for NATA condition classes had the same hierarchy of RCPAQAP bias predictors as presented in Figure 7a (although LD.Bias was removed from the later modelling based on single decision tree results).

The NATA Minor Classes produced a more complex decision tree (Figure 7b), compared to the NATA Condition Classes decision tree (Figure 7c). Both models featured GGT.Bias as the leading predictor, as confirmed by the random forest analysis (Figure 7a), which also showed AST and ALT biases as the second and third ranked predictors respectively. For the decision tree model to predict NATA Classes based on Condition categories (Figure 7c), the GGT node branched into an AST.Bias node only, while the prediction of NATA Minor Classes, grew and extra node from AST to include ALT.Bias in the model, hence the added complexity. This also translated into different NATA Class prediction accuracies, namely;

(7) NATA Minor Prediction: GGT.Bias (≥ 0.0128) + ALT.Bias (≥ 0.0278) = NATA Low (0) Class (68.2%);
(a) GGT.Bias (< 0.0128) = NATA High (1) Class (69.9%) *

* A NATA High (1) prediction of 61.2% was also identified at ALT.Bias (< 0.0278) + AST.Bias (< 0.0628). A 100% prediction NATA (0) Low was found at AST.Bias ≥ 0.0628, but only with 10 total cases available for interrogation (Figure 7b).

(8) NATA Condition Prediction: GGT.Bias (≥ - 0.0117) + AST.Bias (≥ - 0.0448) = NATA Low (0) Class (64.7%);
(a) GGT.Bias (< - 0.0117) = NATA High (1) Class (70.8%) *

* A NATA High (1) prediction of 61.8% was also identified at AST.Bias (< - 0.0448).

The highest accuracies calculated from the Figures 7b decision tree were approximately the same, within the narrow range of 68 - 70% to predict the NATA Minor Classes, whereas the accuracy range from Figure 7c was wider for NATA Condition Class predictions.

Interestingly, the NATA Condition analysis had decision bias thresholds less than 0.0, whereas for the design of decision rules for NATA Minor Classes, the thresholds were > 0.0 bias. Threshold values for both decision tree models ranged between - 0.05 to 0.06 bias.

TABLE 12 - The impact of NATA Class (Minor or Conditions medians) on recursive partitioning models. **(a)** Decision tree modelling and prediction (accuracy calculated via R caret package: 70/30 training - testing split), and **(b)** Random Forest model results reported as out-of-box (OOB) error rates.

NATA Class - Minor or Condition * Reports **(a)**	NATA Minor Class (LFT Model)		NATA Condition Class (LFT Model)	
	Prediction Accuracy (%)	Overall Accuracy (%)	Prediction Accuracy (%)	Overall Accuracy (%)
High (> Median)	67.8	70.2	68.4	73.1
Low (< Median)	73.6		79.1	
NATA Class **(b)**	OOB Error Rate	Overall OOB Error Rate	OOB Error Rate	Overall OOB Error Rate
High (> Median)	24.8	26.6	30.3	27.0
Low (< Median)	28.5		24.2	

(a) Decision tree Model - NATA Minor or Condition Class → RCPAQAP Bias results: GGT + ALT + AST + TBil + TP. Minimum splits = 30 + Complexity Parameter (cp) = 0.025 (Minor); Minimum splits = 30 + cp = 0.030 (Conditions).
(b) Random Forest Model - NATA Minor or Condition Class → RCPAQAP Bias results: GGT + ALT + AST + TBil + LD + TP. 5000 trees interrogated, mtry = 2 (2 variables entered per tree).

* The median number of Minor or Condition reports across all SPLN laboratories was calculated to allow the designation of high (> median) and low (< median) NATA Classes for modelling with RCPAQAP results.

The accuracy of recursive partitioning models was explored further via R caret decision tree training and testing (Table 12a), as well as by a full tuned random forest model (Table 12b), for both NATA Minor and Condition Class modelling.

Considering the Table 12 results, NATA Class prediction varied between 68 - 79%. For the full model (Table 12b), OOB error rates varied between 24 - 30%, which together suggest accuracies between 70 - 76%. The caret (70/30 training - testing split - Table 12a) showed more success for both NATA minor and condition low class prediction, whereas the full model (Table 12b) showed inverted high versus low error rates, depending on whether the model predicted high or low NATA Classes. These results demonstrate the difficulties in analysing these combined NATA and RCPAQAP results.

On considering these results, and after tuning - optimising the models further, additional analyses involving only NATA minor classes and excluding LD.Bias were performed (Table 13). The performance of the caret RF model was superior, with positive and negative predictive values over 70%, a kappa statistic result over 40%, a McNemar's results of p > 0.25, and an overall model accuracy of almost 72% (Table 13b). The PLS model (also Table 13b) was not as impressive, presenting the RF caret optimisation and modelling as the best strategy for the development of integrated NATA and RCPAQAP prediction, in relation to model accuracy and robustness.

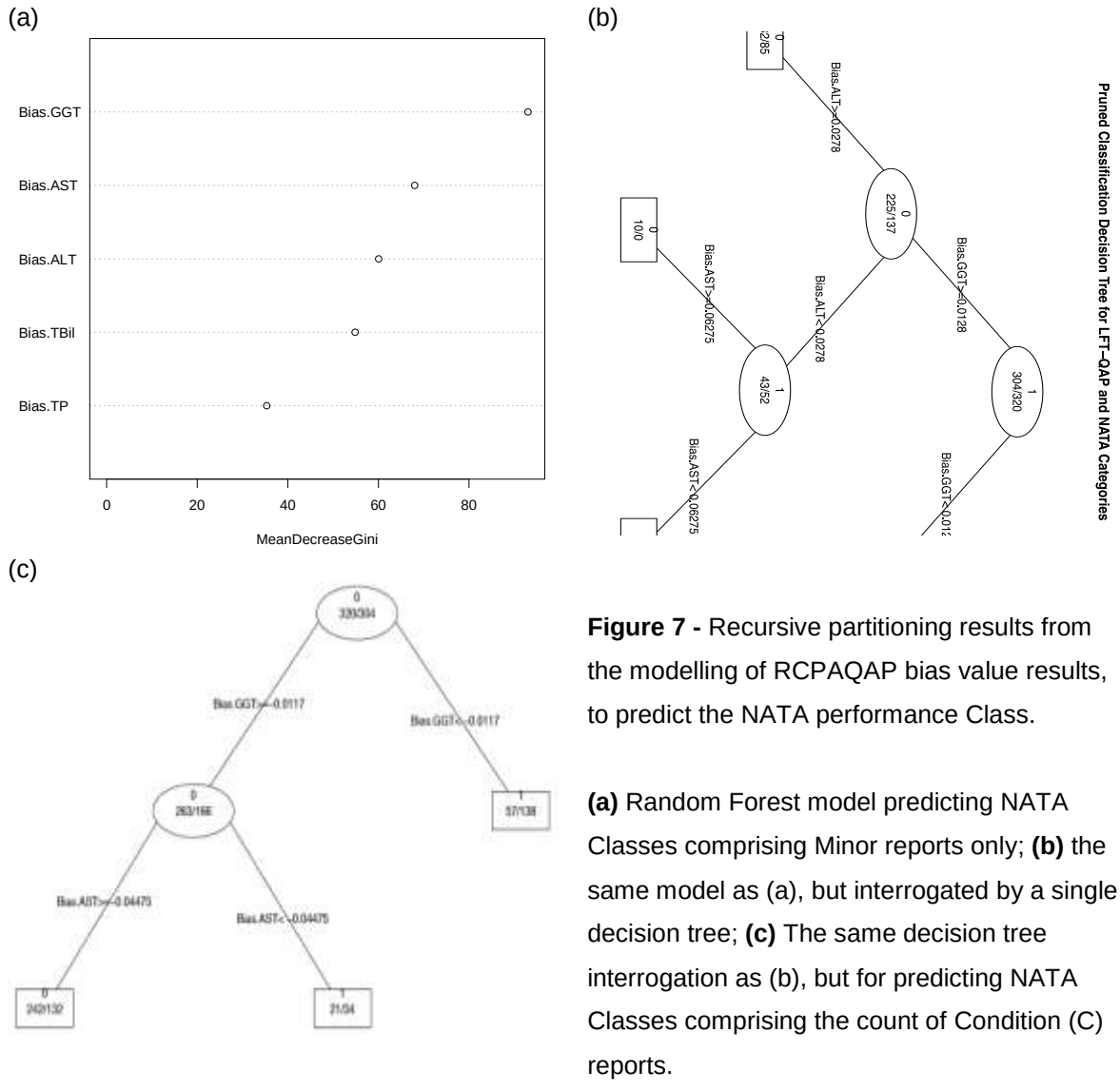


Figure 7 - Recursive partitioning results from the modelling of RCPAQAP bias value results, to predict the NATA performance Class.

(a) Random Forest model predicting NATA Classes comprising Minor reports only; **(b)** the same model as (a), but interrogated by a single decision tree; **(c)** The same decision tree interrogation as (b), but for predicting NATA Classes comprising the count of Condition (C) reports.

Support Vector Machines - Informed by the recursive partitioning results, SVM models were developed that focussed on the prediction of NATA Minor Classes via the top three RCPAQAP predictor variables, namely GGT.Bias, ALT.Bias and AST.Bias (Figure 8).

With GGT and ALT x and y axis variable respectively, AST was introduced as the extra dimension slice at three bias levels. With the AST (bias) slice at - 0.15 (Figure 8a), both

GGT and ALT biases were predicted at > 0.0 for NATA Class 0, with ranges of approximately (~) GGT 0.2 to > 0.4, and ALT at 0.0 to > 0.4. At the neutral slice of 0.0 Bias (exact RCPAQAP target value) (Figure 8b), the GGT range was extended negatively to ~ 0.15, with higher threshold remaining at > 0.4, whereas ALT was condensed to a range of - 0.15 to 0.20. Figure 8c shows that most of the ALT prediction range sat below 0.0, whereas at < - 0.15 GGT predictive range was between 0.0 and 0.3, in relation to NATA low M predictions.

TABLE 13 - Predictive statistical and model tuning parameters of the recursive partitioning analyses presented in Figure 7, which interrogated SPLN LFT (bias) data for the best predictors of NATA class (High or Low NATA Minor condition reports), as assessed via a full randomForest model **(a)**, and caret Random Forest or Partial Least Squares **(b)**.

Features and Results from a full RF Model ***(a)***	**Random Forest (RF) - Full Model**			
	Optimal mtry	**Accuracy**	**Kappa**	**Final Model (OOB Error Rate) - Figure 7a**
Tuned Model (RF)	5	71.96%	ND	**28.04%** High NATA (M) Reports 236/320 = 0.263 Low NATA (M) Reports 213/304 = 0.299
Features and Results from caret Models ***(b)***	**Random Forest (RF) - caret (mtry = 5)**			
	ROC	**Sensitivity**	**Specificity**	**Final Model Accuracy**
Model Tuning	0.785	0.746	0.705	0.7166 (95% CI: 0.646, 0.780)
Final Model Statistics	**McNemar's**	**Sensitivity**	**Specificity**	
	0.272	0.771	0.659	
	Kappa	**PPV**	**NPV**	
	0.431	0.705	0.732	
*Features and Results from PLS Models**	**Partial Least Squares (PLS) - caret**			
	ROC	**Sensitivity**	**Specificity**	**Final Model Accuracy**
Model Tuning	0.717	0.724	0.601	0.668 (95% CI: 0.596, 0.735)
Final Model Statistics	**McNemar's**	**Sensitivity**	**Specificity**	
	0.008	0.792	0.539	
	Kappa	**PPV**	**NPV**	
	0.332	0.644	0.710	

Model interrogated - NATA Minor (High/Low) Class → RCPAQAP Bias results: ALT + AST + TBil + GGT + TP (LD.Bias did not contribute any influence single decision tree results: LD excluded from subsequent models).

Conclusions (SPLN - LFTs) - The combination of recursive partitioning and SVM modelling bring analytic elements that complement the other. The thresholds calculated from single decision trees (Figure 7) were all < 0.10, providing finely tuned predictions that were

complemented by the SVM plots (Figure 8), which illustrated the broader patterns associated with the top (bias) predictors identified by recursive partitioning.

Ultimately the best model for the prediction of NATA outcomes involved a focus on the count of minor reports for the sampled laboratories, using GGT, ALT and AST biases as predictors (with the prominence of GGT and NATA minor reports validating previous results from the earlier pilot study). Therefore, the count of minor reports, which tend to be more numerous, and GGT, are key to the design of an integrative model of laboratory performance.

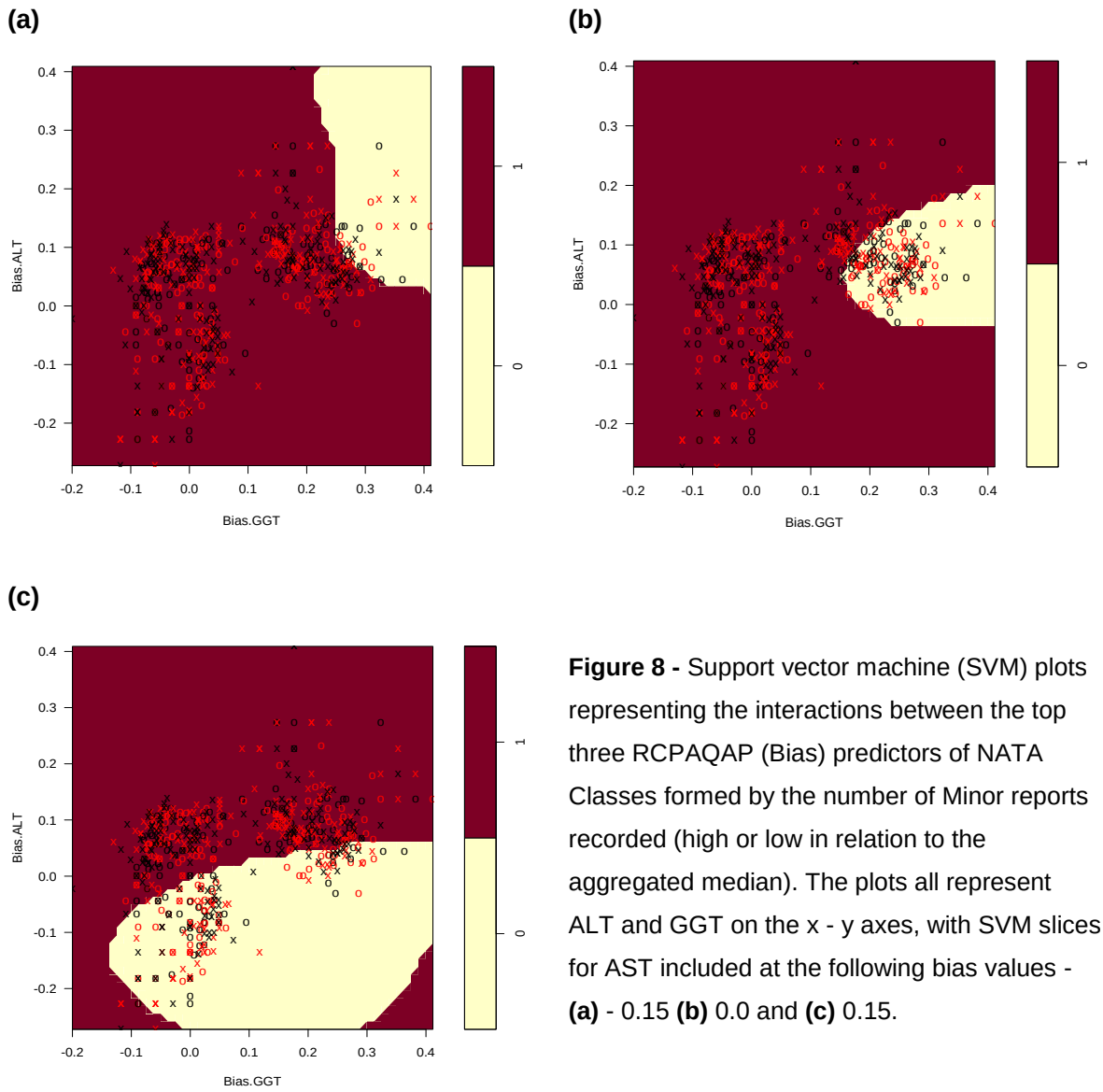


Figure 8 - Support vector machine (SVM) plots representing the interactions between the top three RCPAQAP (Bias) predictors of NATA Classes formed by the number of Minor reports recorded (high or low in relation to the aggregated median). The plots all represent ALT and GGT on the x - y axes, with SVM slices for AST included at the following bias values - **(a)** - 0.15 **(b)** 0.0 and **(c)** 0.15.

(d) Patterns in RCPAQAP-Bias Frequency Distributions: A Possible Explanation for the leading NATA Class Outcome Predictors

Performance Report 4 (July 2019) presented data on the RCPAQAP bias profile of GGT linked to SPLN laboratories. A series of boxplots and error bar plots, representing the RCPAQAP test cycle (1 - 16), demonstrated the pattern of distribution for GGT.Bias when

separated into classes based on higher or lower NATA minor reports. For low NATA M reports, the medians across the 16 RCPAQAP cycles were close to a bias value of 0.20, whereas the high M reports were closer to zero (0.0). This clear distinction in bias patterns was not found for NATA C or C+M Classes, or when the results were separated on the basis of B or G laboratory designation.

As stated above, a key result from this project is the value of GGT (bias) results from the RCPAQAP cycles as a predictor of NATA results represented as classes/categories. Since we have the prediction of two distinct classes (yes/no, high/low), a clear separation in the bias results for GGT, and other important RCPAQAP markers, may help explain their power as NATA Class predictors. To investigate further, RCPAQAP results for GGT and other markers were plotted as histograms, and randomness (from a median of 0.0 bias) was assessed via the non-parametric Runs Test.

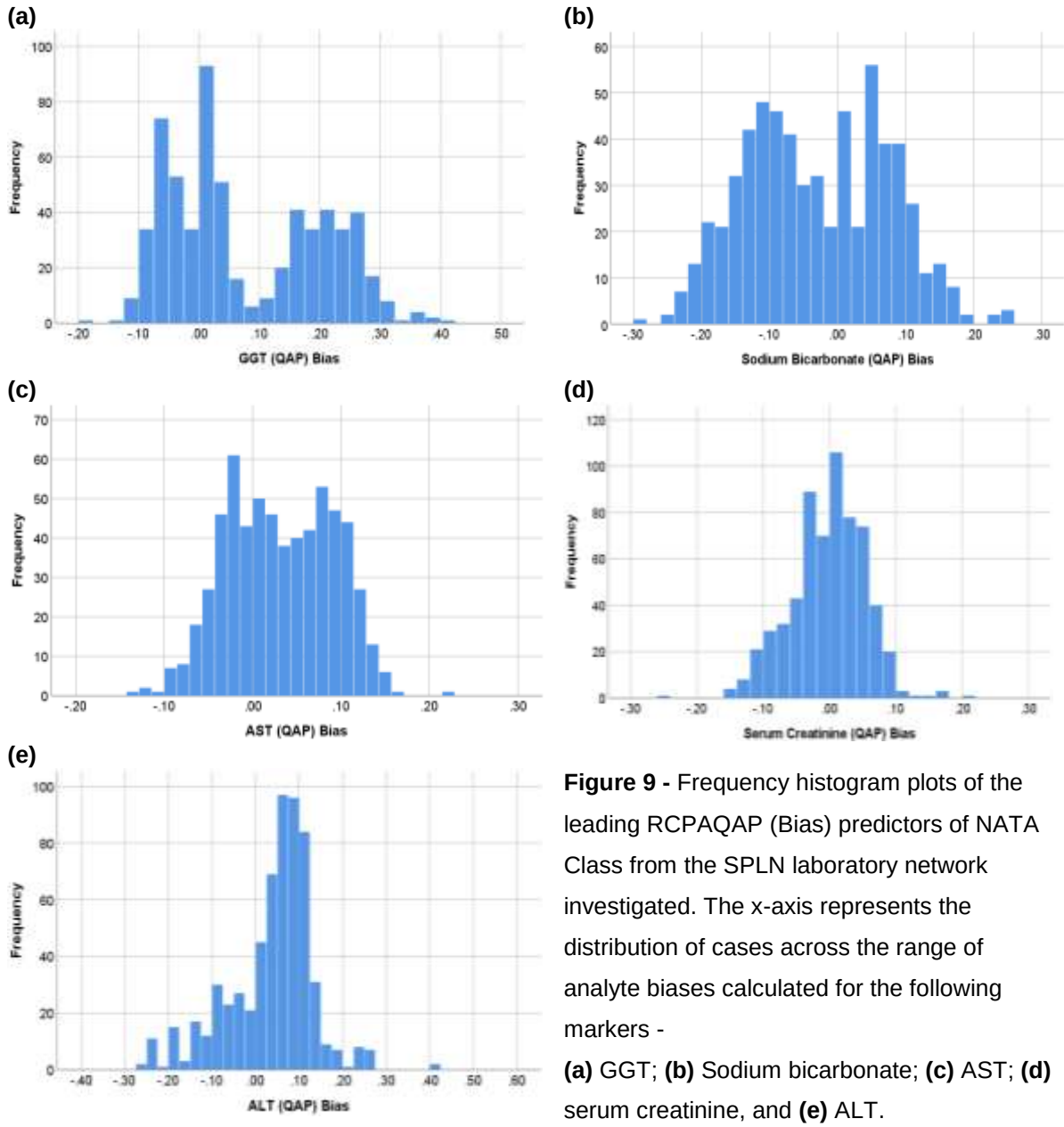


Figure 9 - Frequency histogram plots of the leading RCPAQAP (Bias) predictors of NATA Class from the SPLN laboratory network investigated. The x-axis represents the distribution of cases across the range of analyte biases calculated for the following markers -
(a) GGT; **(b)** Sodium bicarbonate; **(c)** AST; **(d)** serum creatinine, and **(e)** ALT.

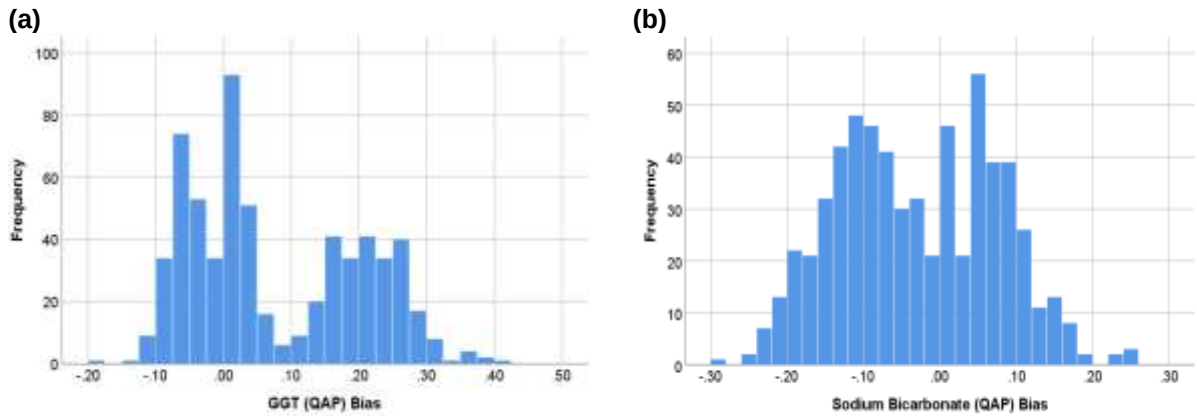


Figure 9 summarises the frequency distribution of leading UEC and LFT (RCPAQAP) predictors of NATA Class, found for SPLN laboratories. As for all analyses presented, the RCPAQAP results were transformed into the relative bias values of individual markers, in relation to the cycle target value. GGT.bias frequency distribution shows a bimodal pattern spanning a bias range of approximately - 0.20 to 0.5, with a separation existing (by virtue of a very low frequency of cases) at approximately 0.10 bias (Figure 9a). The distribution is uneven in relation to a bias of 0.0, with higher peaks from the 0.0 to negative bias range.

This dichotomy in GGT.Bias frequency distribution may explain its predictive power for SPLN laboratories in the context of NATA results. The recursive partitioning and SVM algorithms work by separating complex data into classes, via decision boundaries (partitions) or a separating hyperplane respectively. The GGT distribution, therefore, would render these mechanisms more effective.

While not as pronounced as GGT.Bias, sodium bicarbonate (Figure 9b) and AST (Figure 9c) show a similar bimodal pattern, which again may explain their utility as predictors. Creatinine and ALT bias distributions were skewed, but not bimodal, with broad bias distribution ranges (Figures 9d, e). Other markers with a narrow bias range, e.g. potassium, had low RF rankings as NATA Class predictors.

Statistical analyses determined that the results (Figure 9) were not random (Runs test: $p < 0.001$), and the actual median values for the biases plotted were significantly different to the optimal bias performance of 0.0 (Single sample Wilcoxen Sign Rank test: $p < 0.001$), emphasising the skewed distribution.

Conclusion - In summary, a bimodal frequency distribution with a comparatively broad range in relative bias values is associated with RCPAQAP marker results (e.g. GGT) that have the best performance as NATA predictors for recursive partitioning and SVM modelling. It must be noted that these observations and conclusions relate to SPLN results, and not the findings on PPLN RCPAQAP and NATA integration (results not shown). As explained earlier, the NATA profiles for PPLN and SPLN were different, so not directly comparable.

(e) Point-of-Care Tests

NATA reports and RCPAQAP results specific to point-of-care testing (PoCT) were made available by the SPLN for three regions from their state-wide pathology network. RCPAQAP results associated with troponin I (TnI), blood gas analyses (e.g. pH, pCO_2), UECs, LD and haemoglobin/haematocrit were available, from which representatives were selected for investigation in the context of NATA performance.

(i) NATA Results - Table 14 summarises the NATA feedback and results for each of the three SPLN regions. No conditions (C) were recorded for any region, but two of the three regions recorded minor (M) conditions and all attracted observations (range 14 - 25). The feedback associated with minor reports mention the new *POCelerator* on several occasions, primarily involving problems of correct reporting and maintenance records, and some reference to the need for IT updating. The concerning issue of under-staffing was also noted by NATA assessors, particularly the situation where a senior scientist had supervisory responsibility for 20 - 25 PoCT sites, with unpaid weekend work noted, broadly suggesting a Risk and a Supervision failure. Other issues raised in relation to minor conditions could be linked to inadequate staffing, or overwork by those responsible for PoCT; for example, poor record-keeping, failure to update systems, non-circulation of the PoCT newsletter.

Table 14 - Summary of NATA feedback and reporting on point-of-care testing (PoCT) performance for three regions within the State Pathology Laboratory Network.

NATA Performance	State Pathology Laboratory Network (SPLN) Region		
	Region 1	Region 2	Region 3
Conditions	0	0	0
Minor	6	5	0
Observations	25	14	15
ISO 15189 Clauses Minor Condition Feedback	4.1.2.1, 4.3, 5.3.1.7(i,j), 5.8.1, 5.10.3	4.1.1.3(e), 4.1.2.1, 4.2.1.6, 4.3, 4.9	None
Example Comments (Minor Conditions)	*… transcription of RCPAQAP results; POCelerator functionality; disparate refrigerator temperatures for PoCT reagents (e.g. > 8^{o}C); correct reporting - PoCT maintenance.*	*Inadequate staffing, supervision … PoCT; Documentation of non-conformities; Poor communication of PoCT Newsletter.*	*Assessed and satisfactory.*

(ii) RCPAQAP Results and NATA-RCPAQAP Models - A number of RCPAQAP PoCT markers were available for analysis, with results from twelve time points comprising the full

cycle. Four markers were chosen for further investigation, namely, TnI, pCO_2, serum creatinine and blood glucose. These four RCPAQAP markers were chosen based on the low incidence of missing data (more missingness, less reliable), as well as the extent and diversity of variation within the individual marker range (namely, variation around a median or target value is required for the algorithms employed to separate the response classes under investigation). Identical to the exploration of SPLN and PPLN laboratories, RCPAQAP results from each cycle were calculated as a relative bias value, with a bias score of 0.0 representing exact agreement with the RCPAQAP target value.

Preliminary RF modelling attempted to predict the three individual regions via their NATA profile, but showed poor results with a 98% error in predicting Region 3 (due most likely to a smaller sample size), and single decision trees not including this region in the ultimate model predictions (results not shown). For the next phase of recursive partition modelling, the NATA results for Regions 2 and 3 were combined, which resulted in two NATA classes of similar profile in terms of sample size, and the numbers of minor conditions and observations (Table 14). Therefore, the foundation of machine learning investigations of these data was not based on different profiles of NATA reports (present or absent, above or below the median), and thus cannot answer the questions on definitions of NATA performance. By using SPLN region as the response class allows us simply to determine which of the PoCT markers are most important when assessing the RCPAQAP - NATA quality control systems.

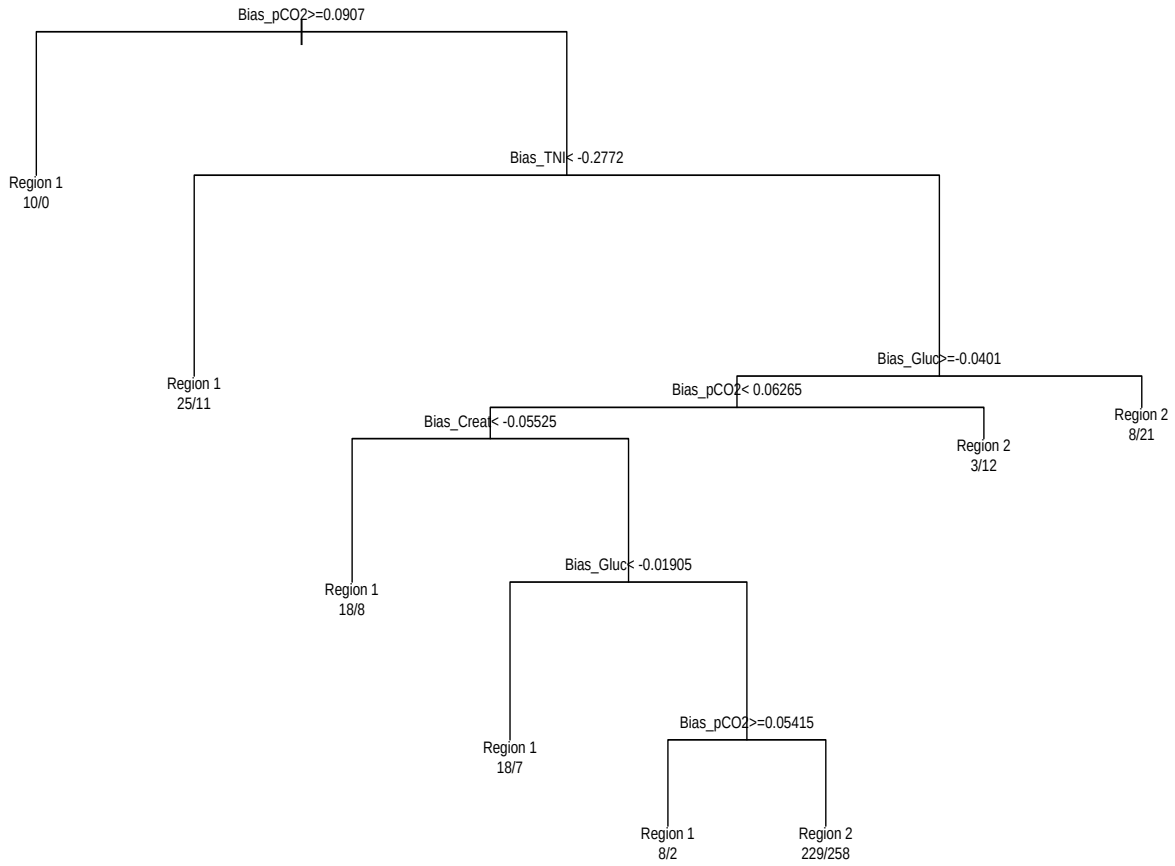


Figure 10 - Single decision tree (recursive partitioning) model of four PoCT RCPAQAP markers (Troponin I, pCO_2, serum creatinine, and blood glucose) in a model using NATA results from three SPLN regions as the response of interest. *Rpart R decision tree model (method = "class", minimum splits per tree = 30, complexity parameter (cp) = 0.016)*.

Figure 10 and Table 15 summarise results from decision trees that examined predictors of SPLN region (region 2 and 3 combined as one region, versus region 1). The resulting decision tree had each RCPAQAP marker represented, with the model assigning variable importance as pCO_2, Glucose, TnI and Creatinine, in descending order (Table 15). Prediction accuracies of 80% were achieved for both Region 1 and Region 2, with both involving pCO_2 at the terminal decision node (Table 15). However, these accuracies were calculated from a small number of cases (12 and 18 respectively), so predictive value may be compromised.

Similar to the examination of RCPAQAP marker bias distribution by histogram (Figure 9), the four PoCT markers investigated by decision tree were similarly plotted as frequency histograms (Figure 11). Creatinine, glucose and pCO_2 showed relatively tight clustering around the perfect bias score of 0.0, with distribution balance generally found on each side of 0.0. This trend was noted also for troponin I (TnI), but with a notable difference in distribution around 0.0 bias (Figure 11d).

TnI had a standard deviation approximately ten-fold greater than the other markers, which can be observed on the x-axis of the TnI plot. On inspection of the raw data, TnI RCPAQAP often showed success at detecting exactly the RCPAQAP target value (hence, a bias score of 0.0), particularly for target values < 0.1, whereas for larger target values (> 1.0) pronounced target misses were seen that reflected bias results of > 0.1. In short, for TnI achieving perfect RCPAQAP results was common for small target values, whereas more variation was seen for higher RCPAQAP targets, although probably not clinically significant. For a future TnI study, it may be worthwhile to run another investigation on only high RCPAQAP target values, while for small concentration TnI results, simple inspection of the data may suffice, with a simple count of how many times the target value was achieved over the RCPAQAP cycle recorded.

TABLE 15 - Summary of the SPLN - Point of Care Test (PoCT) decision tree model presented in Figure 10 (SPLN Region Class prediction: n= 638 - Root and terminal nodes only).

Split	n	loss	y - value	y - probability	
				Correct	Incorrect
Root	638	319	Region 1	0.500	0.500
TnI < - 0.2772	36	11	Region 1	0.694	0.306
pCO_2 < 0.05415	487	229	Region 2	0.530	0.470
Gluc < - 0.0191	25	7	Region 1	0.720	0.280
Gluc < - 0.0401	29	8	Region 2	0.724	0.276
$pCO_2 \geq$ 0.0542	10	2	Region 1	0.800	0.200
$pCO_2 \geq$ 0.0627	15	3	Region 2	0.800	0.200

PoCT markers (TnI, pCO2, creatinine, glucose) represent terminal split values calculated as relative bias scores, where accuracy values are determined by the model.
Variable importance (Weighted value: Arbitrary scale) - pCO_2 (48); Glucose (26); TnI (15); Creatinine (10).

Conclusions (SPLN - PoCT) - Firstly, the investigation of PoCT in the context of NATA inspection results and associated RCPAQAP performance could not be conducted in the same style as for PPLN and SPLN laboratories. The SPLN regions examined had similar NATA report profiles, so RCPAQAP predictions of NATA outcomes could not be examined (for a NATA performance focus, more PoCT results of greater NATA diversity will be required). However, an assessment of the value and “importance” of RCPAQAP PoCT predictors was possible, with some nuances of TnI PoCT revealed for future consideration.

The recursive partitioning and histogram results sit in the context of NATA inspections with no Condition reports, but some Minor reports (Table 14). The feedback from NATA assessors uncovered themes of understaffing and issues with the presumably recent

POCelerator platform. Given these comments, PoCT systems supporting the three regions involved are operating well.

As specific advice from the recursive partitioning analyses, a deeper examination into pCO_2 PoCT may be warranted, as well as the suggested deeper investigation of TnI variation. The volatility of blood gases during collection and measurement most likely explains the consistent recursive partitioning results of pCO_2 as the top-ranked predictor of PoCT region, further suggesting that future studies also involve pH evaluations. As emphasised earlier, the detected variation does not reflect NATA performance, for the reasons previously stated, but differences between SPLN regions indicated the most labile PoCT markers. In the context of this situation, the importance of individual EQA assessment for those sites must be considered since the current NATA monitoring at the regional level may obscure poor performance (11).

Given these overall conclusions, it may be worthwhile to reconsider the quality assurance process pertaining to PoCT; in other words, assess the RCPAQAP performance against the NATA assessment of the entire laboratory network. The benefits of this approach could involve, for example, avoiding the use of out-of-date PoCT cartridges. Presumably this example of operator error may be obviated by consideration of broader laboratory NATA feedback that evaluates management practices (including reagent etc ordering processes).

As explained above, many of the PoCT RCPAQAP results achieved a level of accuracy in obtaining a 0.0 relative bias target that it was obvious via simple inspection of the raw results. In conversations with experts who formerly managed laboratory and PoCT networks, the responsibility for PoCT operations often falls to the “best operators” from the laboratory staff (personal communication), which may explain the higher performance noted anecdotally. Regardless of the operator expertise, complex systems, like a pathology laboratory network, are still prone to management failures and/or simple oversights. Hence the recommendation on assessing the entire laboratory function, not only that pertaining directly to PoCT, may improve the system further.

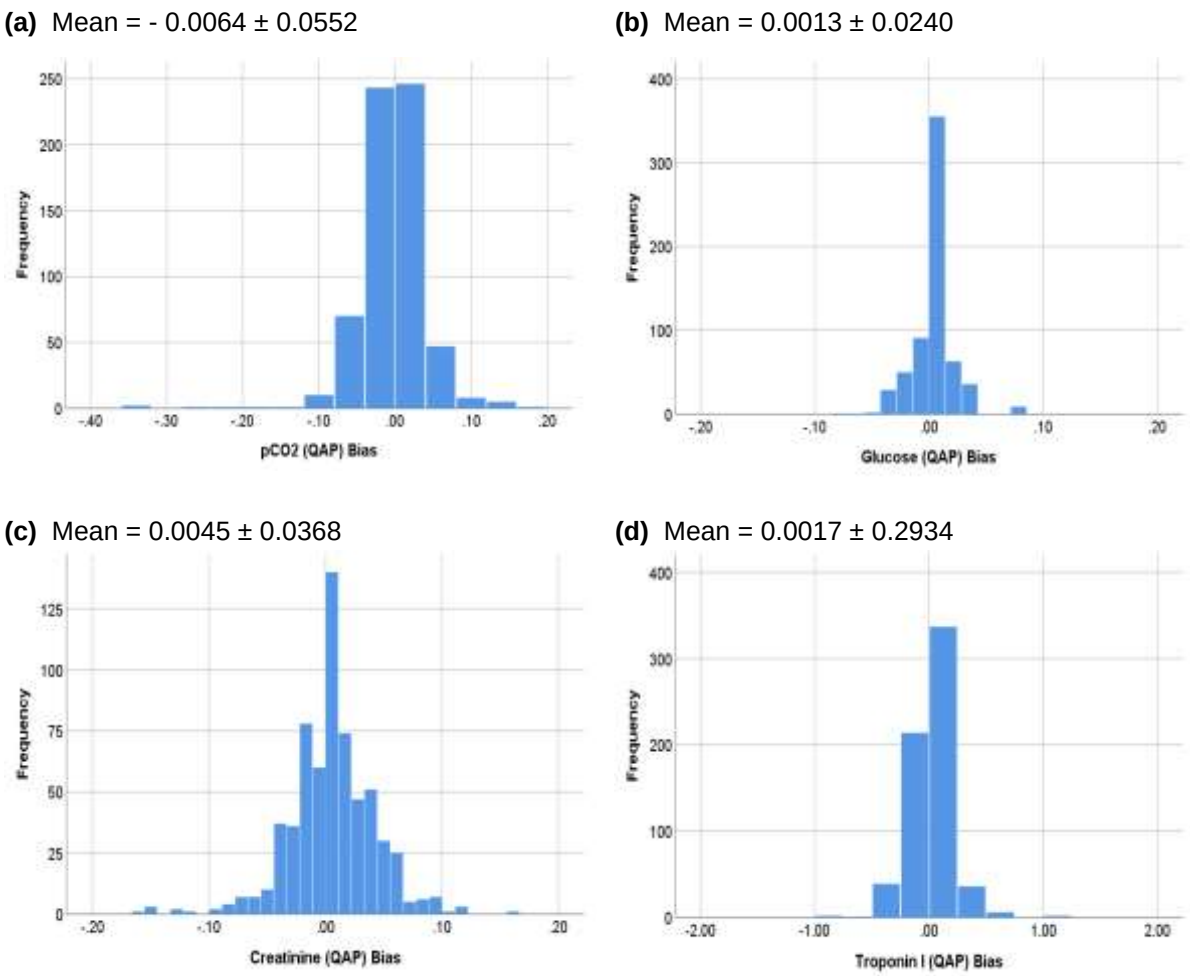


Figure 11 - Frequency histogram plots of selected RCPAQAP (Bias) predictors of NATA Class categories associated with State Region, from the SPLN PoCT network. The x-axis represents the distribution of cases across the range of analyte biases calculated for the following markers - **(a)** pCO_2; **(b)** blood glucose; **(c)** serum creatinine; **(d)** Troponin I. N = 638 RCPAQAP tests of PoCT performance over 12 cycles. Note the wider x-axis scale for Troponin I.

(4) Conclusions and Discussion (Including Benefits Pathology Stakeholders)

The profile of NATA results was very different for PPLN in comparison with SPLN, so a direct juxtaposition between the PPLN models and SPLN was not possible. However, the resulting analyses did draw attention to the potential to adjust NATA - RCPAQAP models according to the number and diversity of NATA reports. In spite of this difference, predictive NATA - RCPAQAP models were developed for serum electrolyte and creatinine markers, and liver function tests profiles, for both pathology networks.

To summarise, the following integrated NATA-RCPAQAP models are proposed for PPLN and SPLN.

To note for these summaries is that all RCPAQAP predictors were transformed to an individual bias result prior to machine learning, modelling, and other analyses (see Methods). C (Conditions), M (Minor) and O (Observations) are the definitions for the categories of NATA reports captured in the NATA Class responses referred to (Methods - PPLN NATA

Classes = Yes or No (i.e. reports present or absent). SPLN NATA Classes = High or Low (i.e. reports above or below the median C or M reports for all laboratories).

Also, of interest is the primacy of GGT as a predictor, including in the context of electrolyte/creatinine RCPAQAP markers. The reason behind the success of GGT (bias) as a RCPAQAP predictor was explored in Results section (e).

Therefore, considering all project results, the following predictive models are proposed:

(a) PPLN

(i) Electrolytes - For the prediction of PPLN laboratories that recorded C, M reports or Observations (CMO Class "Yes") → (RCPAQAP Predictors) = Calcium (Ca^{++}) + Phosphate (PO_4) + gamma (γ) - glutamyl transferase (GGT). Calcium prediction threshold calculated as > or < 0.0017 (relative bias) to predict NATA Class, supported by a GGT range of - 0.05 to - 0.20, with phosphate at a range of approximately - 0.05 to 0.10.

This result must be interpreted with caution as measures of model robustness (McNemar's, Kappa, Rand Index) were generally poor. More data from PPLNs are needed to increase predictive (NATA) model performance.

(ii) LFTs - The first point to note is that GGT was a poor RCPAQAP predictor for PPLN-LFTs, contrary to other results.

For the prediction of PPLN laboratories that recorded M reports (M Class "Yes") → (RCPAQAP Predictors) = Lactate Dehydrogenase (LD) + Aspartate aminotransferase (AST) + Albumin or Alkaline Phosphatase (ALP). Optimal ALP values ranged from - 0.20 to 0.02, supported by AST ranges less than 0.0. The tightest LD ranges (< - 0.02) for NATA M Class prediction were at ALP 0.0 to 0.02 (Figure 4).

Improved McNemar's statistic results (RF) and Rand Index (SVM) suggested robust predictions, with higher prediction accuracies found for RF models of M Class prediction. Kappa values improved, but did not exceed 0.40, indicating poorer performance (Table 8).

(b) SPLN

(i) Electrolytes - As noted, GGT has an interesting synergy with RCPAQAP electrolyte results, which we have reported previously (2), and which was validated again with increased prediction accuracy found when including GGT in the SPLN electrolyte/creatinine model to predict NATA M Classes (Table 11). Again, the best models were found when predicting NATA M Class only (not total C + M). Therefore, the proposed integrated NATA - RCPAQAP model involves GGT, creatinine and sodium bicarbonate. GGT and creatinine bias thresholds are > or < 0.0128 and > or < 0.024 respectively, with broad bicarbonate ranges (< - 0.20 to > 0.20). Therefore, while bicarbonate supports the predictive model, GGT and creatinine

provide the fine-tuned decision thresholds (Figure 6). RF modelling provided the most robust models and highest accuracies (Tables 10 and 11), with increased kappa values and non-significant ($p > 0.05$) McNemar's statistic. Predictive values were > 70%, suggesting a model that after additional calibration could be useful for quality evaluation in the short term.

(ii) LFTs - Predictions of M or C NATA Classes alone showed some value, with respectable prediction accuracies and model robustness revealed (Table 12). Ultimately, the NATA M Class prediction produced a "deeper" decision tree model involving GGT, ALT and AST as the foremost predictor variables. Decision tree thresholds ranged between (+) 0.01 to 0.07 (Figure 7). Further analysis by SVM showed that prediction of low M Classes (coded on the plots as "0"), the optimal AST range was - 0.15 to 0.15, with GGT ranges > 0.15 for the AST range of - 0.15 to 0.0. An AST of 0.15 saw GGT ranges < 0.0, where ALT was mostly less than 0.0. ALT was mostly > 0.0 relative bias at AST values of - 0.15 to 0.0.

Assessment of the model accuracy and robustness by caret again showed best performance through RF (recursive partitioning) with prediction accuracies > 70% (NATA M Class), predictive values > 70%, and reasonable robustness as indicated by Kappa and McNemar's statistics (Table 13). Again, with extra data and optimisation, a useful integrated NATA-RCPAQAP model could be available soon.

(c) Conclusions (Machine Learning Predictive Models)

GGT featured strongly in the predictive modelling, which may be of value to a linking of UEC and LFT RCPAQAP results into a general model of NATA integration. Therefore, a future study is required to develop models with a central GGT role, in combination with the best UEC and LFT RCPAQAP markers identified here. Consideration of routine haematology markers may also be useful to enhance model performance. A study of the full blood count (FBC) markers haemoglobin (Hb), red cell distribution width (RDW), white cell count (WCC) and platelet count (Plt) from the SPLN found that RDW was the best predictor of NATA Class (Results not shown - The PPLN did not provide FBC data, so a full study on PPLN versus SPLN comparisons was not possible on this occasion).

The SPLN also provided RCPAQAP data for special tests, for example drugs and antibiotics, but these had limited sample sizes and were not suitable for machine learning investigations. In general, therefore, routine chemistry and blood markers are the best option for developing integrated quality models to assist laboratories monitor and diagnose performance issues, in collaboration with the RCPAQAP and NATA.

On surveying the pathology quality literature since 2017, only one reference was found that included the concept of computer-based systems to improve ISO 15189 compliance, which was described in the context of training (12). There is discussion in the

literature on the application of AI (Artificial Intelligence) in the context of healthcare broadly (13) that will have some relevance and benefit to pathology.
(NB: The pre-2017 ISO 15189 and associated quality literature was reviewed and reported previously - see references 2 and 3).

As previously concluded (2, 3), the state-of-the-art recommendation for improving laboratory quality was to use root-cause-analysis, with the application of machine learning and AI approaches to quality data not proving popular within the field. This presents an opportunity for the Australian RCPAQAP and NATA to lead the world in the development of the inevitable introduction of AI and Big Data into laboratory medicine. This and previous QUPP - funded projects (2, 3) have demonstrated that machine learning algorithms (as a feature of "AI") can separate complex, integrated data and results into patterns that are able to be interpreted and the results applied to laboratory practice. As a data-driven strategy, more data is required to achieve enhanced predictive, integrated models.

In terms of the data required for RCPAQAP - NATA quality projects, we found two distinct modelling strategies that relied on the diversity and quantity of NATA results. PPLN had fewer NATA reports in comparison with SPLN - for example, only 2 Conditions were reported for PPLN laboratories, while SPLN had 128 Conditions. The number and diversity of NATA reports explains why the SPLN models were more robust, since variation in the data, like for statistical methods, is required to accurately determine the separation of points in a model, and thus provide predictive estimates. To solve the problem of narrower NATA results found for PPLN (only in the context of research models, not practice), more laboratories are required for investigation via the integrated machine learning models presented herein. While the SPLN models featured better predictive accuracies, sensitivities/specificities and positive/negative predictive values, the kappa results suggest that a larger sample size also will be beneficial in relation to model reliability.

The issue of NATA results diversity was also associated with the PoCT study, which had a limited range and variety of data. In spite of this, our investigation demonstrated that pCO_2 and TnI are PoCT markers that need further consideration.

(d) Point-of-Care Tests

As presented above, the PoCT investigations mirrored the situation of the PPLN in that NATA C and M reports were scarce, limiting the extent of an integrated model. Therefore, the recursive partition models of PoCT that explored RCPAQAP relative bias to explain NATA reports could examine only general variation associated with regional differences within the SPLN. As recommended earlier in the report, there are features of pCO_2 and TnI that require further investigation, including the further consideration on PoCT blood gas analysis. Within this scope, consideration of the interpreting RCPAQAP point-of-care results via the overall NATA performance was proposed.

In discussing the ramifications of an alternative approach to NATA reporting on NPAAC guidelines, we need to consider Aim 3(c) - "*Using a small subset of the PoCT sites, use the NPAAC - PoCT Guideline as an assessment guide to assess whether there are associations between failed Clauses, QC or EQA and performance".*

In summary, there were no *"failed Clauses"*, but opportunities for improvement can be recommended based on the wider consideration of the results presented.

In the case of PoCT sites, these are often numerous with many operators using the devices. This complexity makes the conventional model of onsite assessment more problematic in identifying opportunities for targeted improvement. EQA should look at the entire testing process, not just the analytical aspect, and this is particularly true with PoCT where the device is usually reliable, but operator effects, such as competence, out of date cartridges, and problems with reporting, can result in poor patient outcomes (for example, where cartridge QC and EQA sample analyses are performed by experienced staff, in an attempt to conceal the poor performance of other operators). Also, the network is large and involves many small sites performing PoCT with multiple operators, leading to communication and management control difficulties (14).

In this context, the following NPAAC Clauses are most relevant to future revisions that may address the above recommendation (15) -

- *G1.1;*
- *G1.4 (Sub-clause C1.4(i) - sections (a - h);*
- *G1.5 (Sub-clauses C1.5(i) and C1.5(ii);*
- *G5.1 (a - k);*
- *G6.3.*

Full definitions available in Appendix F.

It is interesting to note that reports on the application of PoCT in the Australian context appeared from 2005, focussed, for example, on the challenges of optimising tests for General Practice, and remote and rural settings (16 - 18).

(e) Benefits for Pathology Stakeholders

The benefits of integrated models incapsulating results from the primary quality assessment programmes, NATA and RCPAQAP (EQA), are largely self-evident. Predictive decision thresholds and associated metrics, like those presented here, allow NATA and RCPAQAP results to interact (and *vice versa*), thus offering opportunities for the earlier detection of quality control issues. Also, the recursive partitioning and SVM algorithms can be introduced

into laboratory IT systems to assist with the detection of problems as they arise. Machine learning relies upon the training and testing of data, so as data builds in the system, so too will the capacity to optimise decision support for specific compliance challenges.

The NATA assessment process involves an onsite assessment of the laboratory quality system activities against the ISO 15189 and NPAAC Standards. This is usually a very detailed and expensive process involving both NATA staff and peer assessors. Developing a risk-based approach to the onsite assessment process would be useful for both NATA and the laboratories concerned. Identifying laboratories at risk sooner could allow a more targeted approach by NATA.

In the case of a laboratory network, the use of a model that identifies laboratories within the network where some additional resources could be usefully deployed to improve quality is an obvious advantage to the network, by quickly assisting in the identification of sites where an intervention will reduce error.

(f) Recommendations

Taken as a package, the results presented herein provide direction on integrating NATA inspection results with the outcomes from the RCPAQAP process, and reveals details on what aspects of the models were most predictive of laboratory performance. This was required since the ranking of laboratory RCPAQAP laboratory performance by percentile (and CV%), overall bias or standard deviation did not reflect NATA performance, as summarised by Conditions (C) or Minor conditions (M), at an aggregated level.

Therefore, in designing an integrated quality assessment model (and eventually system) for the timely detection of problems encountered by pathology laboratories that impact result quality, the following are recommended by the results of this project -

1) Convert RCPAQAP data for all analytes and time points to a relative bias value, prior to analysis and modelling (see Methods - not a general laboratory bias value, but bias associated with each RCPAQAP marker at each time point of the cycle);
2) Plot RCPAQAP bias results from each time point over the entire test period as a histogram and examine these plots for a bimodal distribution, with peaks each side of bias 0.0 (e.g., GGT). Plots can be skewed towards a negative or positive relative bias, and this was noted when separating results into B or G laboratories, and C or M NATA Conditions. This pattern was associated with the best QAP predictors of NATA classes (PoCT showed plots approximating a normal distribution, with 0.0 relative bias as the mean);
3) For categories (classes) of NATA performance, use the median of Minor (M) condition counts to determine High or Low classification of laboratories for modelling (namely, counts above the median are assigned the High category, whereas below the median are Low;

4) For aggregated results with fewer C, M or O NATA reports, count all conditions (C or M) and observations (O), and classify laboratory classes according to a presence or absence of these NATA inspection results;
5) Conduct Random Forest (RF) modelling initially to rank RCPAQAP predictors of NATA Class, post tuning and assessment via the R caret package. Use single decision trees (30 trees, complexity parameter (cp) tuning required) and Support Vector Machines (SVM) to detect and define broader patterns of NATA Class prediction, as well as determine decision boundaries/thresholds;
6) Embed these rules into pathology quality control systems, and further tune these parameters via exposure to ongoing NATA and RCPAQAP results.

(g) Future Research

A deeper study is required to assess PoCT integrative modelling via NATA and RCPAQAP results. This study demonstrated that a model can be developed in relation to variation between pathology network regions, but the NATA results were sufficient to adequately integrate the RCAPQAP results with the NATA reports. As recommended, assessing PoCT in relation to overall laboratory function, as captured by NATA, may assist in this regard. In spite of this limitation, the project was successful in identifying as necessary further investigations of blood gases and TnI; for both future studies, considerably more data will be required for interrogation.

Similar to the PoCT results, the investigation of the PPLN was partly constrained by the small number of C and M conditions for the laboratories in the sample provided for investigation. In terms of practice, this is an excellent outcome. For research, however, the reduced variation in the NATA classes lead to predictive integrated models that were not reliable, according to statistical measures of machine learning model robustness, for example, the Kappa statistic and Rand Index. This research challenge can be overcome via the provision of more data from larger laboratory samples.

The SPLN provided NATA profiles of sufficient C and M variation to allow the development of relatively robust models via machine learning algorithms; for example, predictive values of NATA Class prediction greater than (>) 70%. More data will help optimise these predictive models, but they could be trialled for a working laboratory situation in the near future, if desired.

(5) References

1. NATA. National Association of Testing Authorities, Australia. 2017. *www.nata.com.au/nata/*
2. Lidbury BA, Koerbin G, Richardson AM and Badrick T. Development of Predictive Metrics that Allow Early Detection of Poor Laboratory Performance Via Machine-Learning Algorithms to Improve Patient Outcomes and Save Health Resources - Pilot Study and Systematic Review. 2017. Quality Use of Pathology Programme (QUPP). *https://www1.health.gov.au/internet/main/publishing.nsf/Content/qupp-predmetrics*
3. Lidbury BA, Koerbin G, Richardson AM, Badrick T. Integration of ISO 15189 and external quality assurance data to assist the detection of poor laboratory performance in NSW, Australia. Journal of Laboratory and Precision Medicine. 2017;2:*97*.
4. IBM Corp. IBM SPSS Statistics for Windows [Internet]. Armonk, NY: IBM Corp; 2019.
5. R Core Team. R: A language and environment for statistical computing. 2019. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
6. Karatzoglou A, Meyer D, Hornik K. Support Vector Machines in R. J Stat Softw. 2006;15(9).
7. Meyer D, Dimitriadou E, Hornik K, Weingessel A and Leisch F. e1071: Misc Functions of the Dept. of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. 2019.
8. Therneau T and Atkinson B. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. 2019. *https://CRAN.R-project.org/package=rpart*.
9. Liaw A, Wiener M. Classification and Regression by randomForest. 2002. R News 2(3),18 - 22.
10. Kuhn M, Wing J, the R Core Team, *et al*. caret: Classification and Regression Training. 2019. R package version 6.0-84. *https://CRAN.R-project.org/package=caret*
11. Martin C, St John A, Badrick T. Integrating Competence Assessment, Internal Quality Control and External Quality Assurance in a Large PoCT Network. 2020; JPoCT 19(1).
12. Bellini C, Cinci F, Scapellato C, Guerranti R. A computer model for professional competence assessment according to ISO 15189. Clin Chem Lab Med. 2020 Feb 24:/j/cclm.ahead-of-print/cclm-2019-1018/cclm-2019-1018.xml. doi: 10.1515/cclm-2019-1018. Epub ahead of print.
13. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. *Bull World Health Organ*. 2020;98:251-256.
14. Ramsay N, Johnson T, Badrick T. Investigating patient adherence with pathology testing in primary care and how POCT can improve it. 2016. Point of Care 15(4):144-151.
15. Guidelines for Point of Care Testing (PoCT). Department of Health, Canberra Australia (First Edition) 2015. https://www1.health.gov.au/internet/main/publishing.nsf/Content/health-npaac-poctguid
16. Shephard M, Shephard A, Watkinson L, Mazzachi B, Worley P. Design, implementation and results of the quality control program for the Australian government's point of care testing in general practice trial. Ann Clin Biochem. 2009;46:413-419.
17. Shephard MD, Mazzachi BC, Watkinson L, et al. Evaluation of a training program for device operators in the Australian Government's Point of Care Testing in General Practice Trial: issues and implications for rural and remote practices. Rural Remote Health. 2009;9:1189.
18. Bubner TK, Laurence CO, Gialamas A, et al. Effectiveness of point-of-care testing for therapeutic control of chronic conditions: results from the PoCT in General Practice Trial. Med J Aust. 2009;190:624-626.

(6) Appendices

Appendix (A)

Summary - Evaluation of the Activity against the Performance Indicators

What are the key milestones for this project that will identify that you have achieved the objectives of the project?	Milestone Status: Nov 2017 - May 2020
Compile, clean and create data pattern rules from PPLN and SPLN data, as well as troponin TAT studies	Achieved. All data compiled, cleaned and pattern rules designed for B and G laboratories. Troponin data not provided for SPLN or PPLN.
Develop rules for the prediction of PoCT quality utilising data from the SPLN	Achieved. PoCT data from 3 SPLN regions obtained, and an integrated NATA-RCPAQAP PoCT predictive model established.
Publish results and findings in peer-reviewed journals and other literature	Partly achieved. One paper published (Ref 3). 1 - 2 papers in preparation (1 x experimental publication + 1 x review planned). Submission by the end of July 2020 intended.

Appendix (B)

Laboratory rankings in relation to NATA reports - Private Pathology Laboratory Network

(review in conjunction with Table 1)

Appendix (C)

Private Pathology Laboratory Network NATA profile (Conditions and comment details)

Appendix (D)

Laboratory rankings in relation to NATA reports - State Pathology Laboratory Network

(review in conjunction with Table 2)

Appendix (E)

Machine Learning Ensemble investigations, and examples of Support Vector Machine models

Appendix (F)

Summary of NPAAC Point-of-Care-Testing Guidelines.

Appendices B - F are attached separately