acera
Australian Centre of Excellence for Risk Analysis

Report Cover Page

ACERA Project
1101D
Title
Adoption of meaningful performance indicators for quarantine inspection performance.
Author(s) / Address (es)
Andrew Robinson, Australian Centre of Excellence for Risk Analysis Robert Mudford, DAFF Kathleen Quan, DAFF Paul Sorbello, DAFF Matthew Chisholm, DAFF
Material Type and Status
Final Report
Summary

- **Background:** DAFF has adopted a risk-based approach to managing the biosecurity risk of various pathways, including international passengers and mail. During Increased Quarantine Intervention (IQI), introduced in 2001, inspection effectiveness had been used as the primary indicator of inspectorate performance. A risk-based approach to management requires a richer suite of indicators that will better align with DAFF values. ACERA Project 1001i *Performance Indicators* recommended *post-intervention compliance* (PIC) of the pathway as a performance indicator.
- **Overview:** This project focuses on broadening the scope of the indicators, implementing them for the international passengers pathway, and assessing the effect on prioritization of passenger cohorts for further intervention.
- **Outcomes:**
  - The recommended indicators are:
    * Before intervention compliance (BIC),
    * Post-intervention compliance (PIC),
    * Non-compliance effectiveness (NCE), and
    * Hit rate (HR).

    These indicators are simple and robust measures of performance, accounting for compliance and inspectorate performance before and after arrival.
  - Three of the indicators can be computed with existing data collections, namely BIC, PIC and NCE. HR can be computed for some sub-pathways, but better tracking information is needed, that is, information about all the intervention steps that the passengers have followed.

- **Outcomes (ctd):**
  - The data prior to June 2012 were not sufficiently detailed. Collection categories have been amended to enable calculation of all indicators at the desired granularity.
  - No substantial implications are anticipated for the profiling methodology as a result of adopting these performance measures. With profiling, the categories with the highest approach rate are targeted. The approach remains the same under the new performance measures.
  - The performance measures can be used to produce new standard reports for monitoring performance. These reports are control charts—as used for statistical process control—but are tailored to DAFF's operational environment. Importantly, they include confidence intervals to show the uncertainty in each performance indicator. Examples are provided in Chapter 6.

- **Recommendations:**
  - The reported performance indicators should be used to assess how appropriately and how well the inspectorate performs, with PIC as the key indicator (p 23).
  - Profiles for international passengers and mail articles should still be based on the approach rate (p 24).
  - Performance indicators should be reported with confidence intervals wherever possible, to enable accurate assessment of the quality of the available information (p 14).
  - The nominal coverage of the confidence intervals should be no less than 90% (p 14).
  - DAFF should determine what would be the effect upon the statistical qualities of the performance indicators of using a sampling approach to counting Incoming Passenger Cards (IPCs) instead of counting all of them (p 21).
  - DAFF should undertake a further study to determine when and how to cluster small cohorts, and what is the effect upon profiling of that clustering, and what other options—for example, empirical Bayes estimates—might be available for handling small cohorts (p 22).
  - DAFF should consider whether the cutoff for high-risk cohorts should be the mean approach rate or some higher confidence interval (or, before-intervention compliance, BIC, or lower interval). The mean is the best indicator of compliance, but the higher confidence interval acknowledges that ignorance is a source of risk (p 32).
  - Leakage surveys need to be representative in order to reduce uncertainty about the compliance of individual cohorts. DAFF should investigate how to assess and report the representativeness of the leakage survey (p 33).
  - DAFF should review the choice of performance indicators and the data collection procedures within one year (p 23).

ACERA Use Only	Received By:	Date:
	ACERA / AMSI SAC Approval:	Date:
	DAFF Endorsement: ( ) Yes ( ) No	Date:

Adoption of meaningful performance indicators for quarantine inspection performance.

ACERA 1101D

Andrew Robinson, Australian Centre of Excellence for Risk Analysis
Robert Mudford, DAFF
Kathleen Quan, DAFF
Paul Sorbello, DAFF
Matthew Chisholm, DAFF

May 12, 2013

Acknowledgments

This report is a product of the Australian Centre of Excellence for Risk Analysis (ACERA). In preparing this report, the authors acknowledge the financial and other support provided by the Department of Agriculture, Fisheries and Forestry (DAFF), the University of Melbourne, Australian Mathematical Sciences Institute (AMSI) and Australian Research Centre for Urban Ecology (ARCUE).

We are also grateful to Mark Debeljakovic, Tony Arthur, Rob Cannon, Greg Hood, and Stephanie Quispes-Garay for very useful review suggestions that improved the report markedly.

Contents

List of Tables

List of Figures

Table 2: Table of definitions of terms used throughout the report.

Term	Definition
Compliant	A *compliant* passenger or mail article is a passenger or mail article that is compliant with all biosecurity regulations.
Effectiveness	*Effectiveness* is taken to mean the quality of intervention, usually the quality of inspection, and is commonly defined as the probability that existing contamination will be detected and rectified. That is, if a unit is contaminated, the effectiveness of inspection is the probability that the contamination will be detected if the unit is inspected.
Efficiency	The *efficiency* reflects the amount of effort that is needed to intercept a contaminated unit. Efficiency is usually reported for screening interventions.
Inspection	*Inspection* refers to the manual examination of a passenger's person or one or more personal effects, or a mail article.
Intervention	*Intervention* is a collective label for different kinds of biosecurity actions, such as examination of the Incoming Passenger Card (IPC), screening based on X-ray or detector dogs, and inspection.
Leakage	*Leakage* is the amount of undetected biosecurity risk material that passes through an intervention point. Leakage can be reported as a rate or as a count.
Non-compliant	A *non-compliant* unit is a unit that is not compliant with at least some biosecurity regulations.
Pathway	The *pathway* is defined as a collection of activities that culminate in the arrival to Australia of a set of alike inspection units. Pathways can be subdivided to reflect management constraints or to enable focusing inspection resources on sub-pathways that are thought to be most risky. Examples are: the arrival of passengers, or the arrival of passengers from a particular country.
Processing	*Processing* is used as a synonym for intervention to replace the clumsy construction "intervened with" with "processed".
Screening	*Screening* refers to the capture and use of information to determine follow-up activity for a passenger or mail article. Examples include profiling based on examination of the IPC or passenger interview, and releasing or referring passengers for manual inspection based on the outcome of X-ray or detector dog intervention.
Unit	The *unit* or *inspection unit* will be the entity that is singled out by the pathway manager for intervention. Examples include an air passenger and a mail article. The definition of the inspection unit is subjective, and usually based on operational convenience.
Volume	The *volume* is the number of units on the pathway.

1

Executive summary

1.1 Findings

1. This report is the first deliverable for ACERA project 1101D, *Adoption of meaningful performance indicators for quarantine inspection performance*. The objective of the project is to show how to measure, use and interpret the performance indicators developed in ACERA Project 1001i *Performance Indicators*, and to investigate any deficiencies in recording systems that prevent their calculation in the passenger and mail pathways. The indicators are:

   - Before intervention compliance (BIC),
   - Post-intervention compliance (PIC),
   - Non-compliance effectiveness (NCE), and
   - Hit rate (HR).

   These indicators are simple and robust measures of performance, accounting for compliance and inspectorate performance before and after arrival.

2. Three of the indicators can be computed with existing data collections, namely BIC, PIC and NCE. HR can be computed for some sub-pathways, but better tracking information is needed, that is, information about all the intervention steps that the passengers have followed.

3. The data prior to June 2012 were not sufficiently detailed. The collection categories have since been amended to enable calculation of all indicators at the desired granularity.

4. No substantial implications are anticipated for the profiling methodology as a result of adopting these performance measures. With profiling, the categories with the highest approach rate are targeted. The outcome remains the same under the new performance measures.

5. The performance measures can be used to produce new standard reports for monitoring performance. These reports are control charts—as used for statistical process control—but are tailored to DAFF's operational environment. Importantly, they include confidence intervals to show the uncertainty in each performance indicator. Examples are provided in Chapter 6.

1.2 Recommendations

As a result of these findings, we make the following recommendations:

1. The proposed performance indicators (BIC, PIC, NCE, HR) should be used to assess how appropriately the inspectorate performs as well as how well it performs, with PIC as the key indicator (p 23).
2. Profiles for international passengers and mail articles should still be based on the approach rate (p 24).
3. Performance indicators should be reported with confidence intervals wherever possible, so that the manager can accurately assess the quality of the available information (p 14).
4. The nominal coverage of the confidence intervals should be no less than 90% (p 14).
5. DAFF should determine what would be the effect upon the statistical qualities of the performance indicators of using a sampling approach to counting Incoming Passenger Cards (IPCs) instead of counting all of them (p 21).
6. DAFF should undertake a further study to determine when and how to cluster small cohorts, and what is the effect upon profiling of that clustering, and what other options—for example, empirical Bayes estimates—might be available for handling small cohorts (p 22).
7. DAFF should consider whether the cutoff for targeting high-risk cohorts should be the mean approach rate or some higher confidence interval (or, BIC or lower interval). The mean is the best indicator of compliance, but the higher confidence interval acknowledges that ignorance is a source of risk (p 32).
8. Leakage surveys need to be representative in order to reduce uncertainty about the compliance of individual cohorts. DAFF should investigate how to assess and report the representativeness of the leakage survey (p 33).
9. DAFF should review the choice of performance indicators and the data collection procedures within one year (p 23).

2

Introduction

The objective of this report is to show how to measure, use, and interpret the performance indicators developed in ACERA Report 1001i1 (**?**). The examples in this report will focus on two pathways: international passengers and mail. Unless otherwise stated, material herein will cover both pathways. Further reports for this project will guide the use of these indicators in other pathways.

The report is structured as follows. The next chapter introduces and defines the suite of performance indicators. Chapter 4 describes the data collection that is needed to calculate the performance indicators, and provides equations to calculate them. Chapter 5 describes the current approach to profiling (focusing on international passengers) and also how that approach should change using the new performance indicators. Chapter 6 provides examples of tables and graphs that can be used to summarize various aspects of performance.

3

Performance Indicators

3.1 Introduction

Following ACERA Report 1001i1 (**?**), the Passengers and Mail Branch elected to implement the following performance indicators:

- Before intervention compliance (BIC),
- Post-intervention compliance (PIC),
- Non-compliance effectiveness (NCE), and
- Hit rate (HR).

We briefly describe the arrival process for international passengers with a focus on the elements that are relevant to this report. A more detailed but somewhat dated description can be found in **?**, from which some of the following text has been copied. We do not cover the mail pathway in this section.

Air passengers usually arrive at one of eight major international airports. After arrival, the passengers are processed by Customs officials at the Entry Control Point (ECP, also called the Primary Line). The Customs officials examine the passenger's passport and Incoming Passenger Card (IPC), ask any follow-up questions, and identify quarantine declarations.[1]

Passengers then move into the hall, which is the area that contains the baggage carousels. DAFF risk assessment officers (RAO) operate in the hall: Hall RAOs (HRAO). HRAOs are generally positioned to interview the passengers as they move from ECP towards or around the carousel. For each interviewed passenger, the HRAO may mark the IPC by pen or by specially-made stamps to indicate a recommended processing method. Generally speaking the HRAO will mark the IPC for one of three outcomes: Release, Screening (by X-ray or detector dog unit, DDU, dogs), or manual inspection. The outcome will depend on whether the passenger has declared any items and also the outcome of the assessment of the passenger's profile made by the HRAO.

The passengers proceed to the Marshal point, at which they are directed to further Customs or DAFF intervention depending on the IPC codes. Passengers that are directed to DAFF screening may then undergo manual inspection based on the outcome of the screening. Manual inspection involves opening one or more items.

All passengers are subject to a leakage survey after DAFF intervention. The leakage survey is a random manual inspection of all unopened bags from the passenger along with recording of some passenger details. The bag selected for inspection must be one that has not already been manually inspected. The leakage survey is assumed to be 100% effective. The outcomes of the inspection and the data capture are both used for computing the performance indicators. Approximately 80,000 leakage survey samples are taken every year across all regions. Some

[1] This protocol varies modestly when the passenger elects to be processed by the new Smartgate facility.

jurisdictions (eg., USDA APHIS) use a different kind of leakage estimate, which has also been considered by DAFF in the past; this point is discussed in Section 4.6.

There are two main considerations for data collection: first, are the appropriate measurements being taken for the desired outcome? And second, are they being taken at the right level to be useful? For example, it is important to know the volume of passengers on the pathway. This is an example of the right measurement. It is also important to know the volume of non-declarant passengers that undergo manual inspection having been directed there by an X-ray operator. This is an example of the right measurement at the right level. Existing data-capture protocols in the International Passengers and Mail pathways collect the right measurements. The measurements are collected at the right level for most of the four proposed indices; changes have recently been made to align the data capture with the requirements for computing Hit Rate.

3.2 Definitions

3.2.1 Before-Intervention Compliance (BIC)

The BIC of a pathway is defined as the proportion of arriving units that are compliant with biosecurity regulations. The BIC is simply one minus the approach rate, which is the common way that DAFF measures the inherent riskiness of a pathway. We advocate reporting BIC in place of the approach rate in order to provide greater compatibility with PIC (qv). However, we believe that profiling should still be done using the approach rate.

3.2.2 Post-Intervention Compliance (PIC)

The PIC of a pathway is defined as the proportion of units that are compliant after all DAFF intervention has been performed. PIC was recommended as an indicator of inspectorate performance by **?**, and is measured and publicly reported for various pathways by USDA APHIS and NZ MPI. The PIC is one minus the leakage.

3.2.3 Non-Compliance Effectiveness (NCE)

The NCE is defined as follows. When a non-compliant unit is inspected, the NCE is the probability that the non-compliance is detected. When a non-compliant unit is screened, the NCE is the probability that the unit is referred for further intervention. NCE can be thought of as the quality of individual interventions, in terms of how successful those interventions are in detecting non-compliance. If the effectiveness is 1, then all screened non-compliant units are referred, or all inspected non-compliant units are intercepted. If the effectiveness is 0, then none of the non-compliant units are referred or intercepted.

3.2.4 Hit Rate (HR)

The HR is a measure of efficiency of screening, and is defined as the proportion of units referred for inspection that are non-compliant. HR is calculated for screening interventions, because screening is used to reduce the number of compliant units that are referred for inspection, that is, it is used to make inspection systems more efficient.

A difficulty with computing HR for screening interventions is that confirmation of the detections occurs downstream, as it were, by inspection. This inspection will likely not detect perfectly, hence the HR of the screening step must be adjusted to account for the units that were correctly profiled but were not detected by the imperfect detection.

As an extreme example, imagine that 20 units from 100 were correctly referred to inspection by a screening procedure, but the subsequent inspection detected only 10 of them, and the

leakage survey estimated that the inspection was only 50% effective. In this case the HR of the screening is estimated as 1.

Presently the data holdings are insufficient for computing HR. The reason for this is that the pathway that a passenger has taken to arrive at manual inspection has not historically been recorded. For example, the passenger may have been directed to manual inspection based on screening by X-ray, or because they declared that they were carrying biosecurity risk material, but neither reason has been recorded. These issues were addressed in the data changes of June 2012.

HR is an important and intuitively attractive performance indicator, but it cannot be used in isolation. As noted by **?**, inspections are performed for more than one reason: intercepting non-compliance is important, but so is keeping up-to-date information on the relative risks of all the pathways.

If no profiling is performed then intervention decisions will be random and the HR will be the same as the approach rate of non-compliant passengers or mail articles, so long as inspection is 100% effective. If profiling is performed then the Hit Rate will ideally be higher than the approach rate, proportional to the specificity of the profiling.

3.3 Example of Indicators

An explanatory diagram is presented in Figure 3.1. The 20 incoming passengers are *screened*, and eight are sent for inspection. Of the eight, one is non-compliant (so HR= $1/8 = 12.5\%$). Twelve passengers are released. Four of the twelve are then inspected in the leakage survey, and one of the four is non-compliant ($1/4 = 25\%$). We therefore estimate that 25% of the remaining eight passengers were non-compliant, which is two. Hence we estimate that $(20 - 2)/20 = 90\%$ of the departing passengers are compliant — this is PIC. We detected two non-compliant passengers and we estimated that there were two more undetected, so our estimated approach rate is $(2 + 2)/20 = 20\%$, and the estimated BIC is $100 - 20 = 80\%$. Finally our inspection detected one of the four estimated non-compliant passengers, so the joint screening and inspection NCE is $1/4 = 25\%$. Our leakage survey caught another, so the total intervention NCE, NCE* in the figure, is $(1 + 1)/4 = 50\%$. Here, the inspection is assumed to be fully effective.

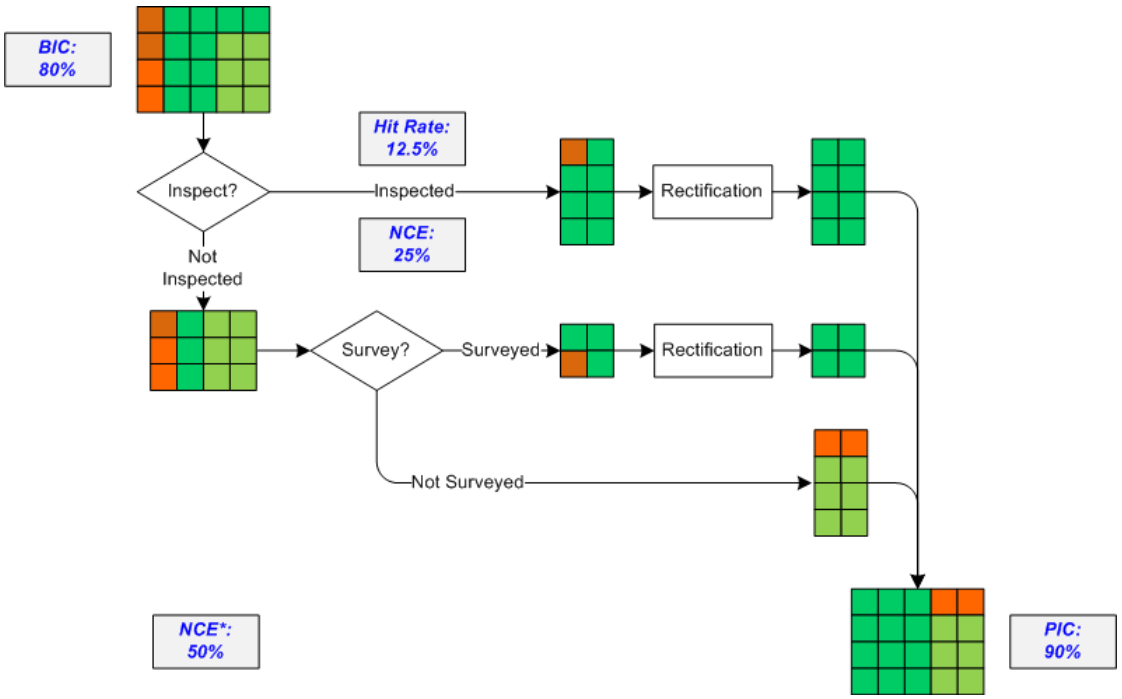

Figure 3.1: Example application of performance indicators to intervention data. See text for explanation. Dark green units are compliant units that are inspected. Light green units are compliant but not inspected. Orange units are non-compliant uninspected units and brown units are non-compliant, inspected units that are subsequently rectified.

4

Data Collection and Indicator Estimation

This chapter presents a collection of protocols for the collection of the data necessary to compute the four performance indicators outlined in Chapter 3.

Performance indicators are statistics, often computed from incomplete information. In the current setting, for example, the indicators rely on estimates of leakage taken from a leakage survey. The leakage estimates have statistical uncertainty because only a sample of the passengers is captured by the leakage survey. The statistical uncertainty of the indicators is best expressed using a confidence interval estimate, also called an interval estimate. We recommend that the performance indicators be accompanied by interval estimates, and we provide guidelines for doing so in the following report. Interval estimates must be accompanied by a statement of coverage, which can be thought of loosely as a statement of the confidence that the interval covers the true value. Coverage of 95% is commonly used for scientific work, but may be higher than is needed for reporting purposes. We recommend that at intervals of at least 90% coverage be reported.

We focus now on the international passenger pathway. The reader should understand that from here on, when we refer to passengers, the text is also relevant for mail articles.

4.1 Indicator Estimates

The essential measures are:

- v, the *volume*, which is the number of units on the pathway;
- i, the number of units *inspected* after screening;
- b, the number of inspected units that were *non-compliant*;
- n, the number of units *processed* in the leakage survey; and
- y, the number of units that were found to be *non-compliant* in the leakage survey.

The following development, which largely follows **?**, is presented as an example given the inspection setup outlined in Figure 4.1. Formal implementation of the indicators will likely deviate from this presentation because of the complexity of the passenger inspection system.

Note that detailed data collection is necessary in order to compute the suite of performance indicators across all border activities for passengers. Specifically, Figure 4.1 simplifies across a complex array of possible pathways, including several layers of screening, and declarant vs. non-declarant passengers. It will be important to ensure that passenger activity counts are sufficiently fine-grained to capture the statistics of interest. Data collection has been in place at the international airports since June 2012.

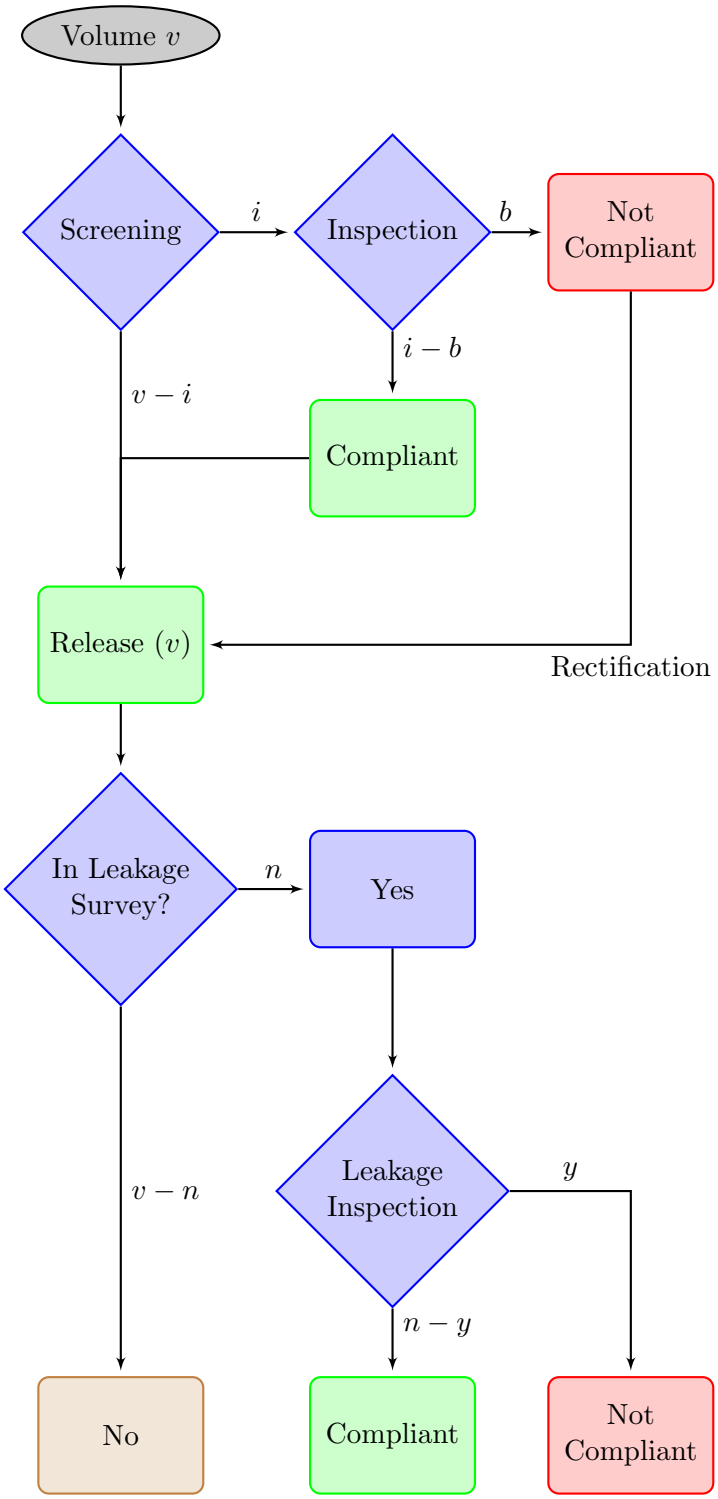

Figure 4.1: Flow chart for sampled intervention of pathway with leakage survey. *Rectification* means that the biosecurity risk material is confiscated from the passenger, and the passenger is then assumed to be compliant. The leakage survey records whether the unit was released or inspected after screening.

4.1.1 Leakage Count and Approach Rate

There are two sources of leakage to consider: one from the inspected units, and one from the released units. The leakage survey records whether the unit was released or inspected after screening. If the leakage survey also records whether or not the unit has been rectified after inspection, then the leakage from rectification can also be computed, but we do not include this detail. Here we use the subscript i to refer to the leakage survey results for inspected units and r to refer to the leakage survey results for the released units. The leakage count from an intervention, l, is estimated using the proportion of units surveyed that were found to be non-compliant, scaled by the number of units, as follows:

$$\hat{l} = i \times \frac{y_i}{n_i} + (v - i) \times \frac{y_r}{n_r} \tag{4.1}$$

Note that leakage is possible even after a passenger has been found to be non-compliant and has been rectified, for example if the passenger has more than one item of biosecurity high-risk material.

The adjusted Wald interval estimate from **?** follows. Let $\hat{p}_i = (y_i + 1)/(n_i + 2)$ and $\hat{p}_r = (y_r + 1)/(n_r + 2)$, and

$$s_{\hat{l}} = \sqrt{i^2 \times \frac{\hat{p}_i \times (1 - \hat{p}_i)}{n_i + 2} + (v - i)^2 \times \frac{\hat{p}_r \times (1 - \hat{p}_r)}{n_r + 2}} \tag{4.2}$$

Then the interval estimate for the leakage count is

$$\hat{l}_I = i \times \hat{p}_i + (v - i) \times \hat{p}_r \pm 1.96 \times s_{\hat{l}} \tag{4.3}$$

The estimated approach count $\hat{a}$ for a pathway is the sum of the detected non-compliant units (b) and estimated undetected non-compliant units ($\hat{l}$).

$$\hat{a} = b + \hat{l} \tag{4.4}$$

The interval estimate for the approach count is

$$\hat{a}_I = b + \hat{l}_I \tag{4.5}$$

Note that b is known exactly so does not contribute to the uncertainty of the interval estimate.

4.1.2 BIC

The BIC is estimated as the difference between the volume and the approach count, scaled by the volume.

$$BIC = \frac{v - \hat{a}}{v} \tag{4.6}$$

An interval estimate for BIC can be obtained by replacing $\hat{a}$ with $\hat{a}_I$ in the equation above, when v is known exactly.

The case for when v is estimated, such as when passenger sub-pathway volumes are computed using raking (**?**), is still under examination. It seems reasonable to assume that the uncertainty of $\hat{v}$ is minor compared with the uncertainty in $\hat{a}$, and therefore that it can be ignored.

4.1.3 PIC

To get the PIC we need to estimate the number of non-compliant units leaked after intervention. The estimated pathway leakage count is the difference between the estimated intervention leakage count and the number of non-compliant units intercepted in the leakage survey.

$$\hat{L} = \hat{l} - y_i - y_r \tag{4.7}$$

An interval estimate for $\hat{L}$ can be obtained by replacing $\hat{l}$ with $\hat{l}_I$.

The reduction of the leakage count by the items intercepted in the leakage survey has been the subject of some discussion (**?**). Here we consider that the leakage survey is an integral component of DAFF intervention, and as such, contaminated items that are intercepted by the leakage survey should not be considered to have leaked.

The PIC is the difference between this quantity and the volume, scaled by the volume.

$$PIC = \frac{v - \hat{L}}{v} \tag{4.8}$$

An interval estimate for PIC can be obtained by replacing $\hat{L}$ with $\hat{L}_I$ when v is known exactly. As above, the case when v is estimated, for example from raking, is under examination, and it seems reasonable to ignore uncertainty in $\hat{v}$.

4.1.4 NCE

The estimated NCE for inspection methods (as opposed to screening) is the ratio of the number of non-compliant units detected during intervention and the estimated approach count.

$$NCE = \frac{b}{\hat{a}} \tag{4.9}$$

An interval estimate for NCE can be obtained by replacing $\hat{a}$ with $\hat{a}_I$, when b is known exactly.

For some intervention methods, the number detected is not known exactly because it is estimated using a further intervention. An example of such an intervention is screening by X-ray. When X-ray screening detects non-compliance, a manual inspection is undertaken. The manual inspection may be imperfect, so the non-compliance detected by the X-ray may be missed by manual inspection. Then,

$$NCE = \frac{b + \hat{l}_i}{\hat{a}} \tag{4.10}$$

where

$$l_i = i \times y_i / n_i \tag{4.11}$$

For this case, the calculation of confidence intervals is more complicated, and is described in Appendix B.

The effectiveness of a screening procedure can also be computed if a leakage survey is taken of units that are released by screening. This is useful as it provides insight as to whether it would be better to improve screening or inspection capability. Effectiveness of screening can be computed as

$$NCE = \frac{1 - l_r}{1 - BIC} \tag{4.12}$$

This quantity gives a theoretical maximum for effectiveness if inspection were 100% effective.

4.1.5 HR

The HR is a measure of efficiency that is calculated for screening methods. It is defined as the count of non-compliant units referred for inspection divided by the count of units referred for inspection.

$$HR = \frac{b + \hat{l}_i}{i} \tag{4.13}$$

Here, $\hat{l}_i$ is the number of non-compliant units referred but not detected in inspection (estimated from the leakage survey), and i is the number of units referred. A reasonable interval estimate for HR follows. First,

$$s_{\hat{l}_i} = \sqrt{i^2 \times \frac{(y_i + 2) \times (n_i - y_i + 2)}{n_i \times (n_i + 4)^2}} \tag{4.14}$$

then

$$\hat{l}_{i_I} = i \times \frac{y_i + 2}{n_i + 4} \pm 1.96 \times s_{\hat{l}_i} \tag{4.15}$$

following **?**. Strictly speaking the adjustments of 2 and 4 to y_i and n_i and the 1.96 are related to the confidence limits: $y + z/2$, $n + z$ and z where z is the 0.975 normal distribution quantile.

$$HR_I = \frac{b + \hat{l}_{i_I}}{i} \tag{4.16}$$

Note that the interval estimate for HR is not symmetric around the point estimate of HR. Also note that 'Assess and Release' passengers are not further screened or inspected so they are not included in the Hit Rate. The intervals proposed by **?** are precursors of the adjusted Wald intervals proposed by **?**.

4.2 Example

This section provides an example of the calculation of the performance indicators for a pathway, using the open-source statistical environment R. The examples have been cut from R output, and look something like this:

```
> (l_hat = i * (y_i / n_i) + (v - i) * (y_r / n_r))

[1] 66.66667
```

The angle bracket `>` is R's way of asking for something to do, as is the `+` sign. Here, we have asked R to calculate the inspection level leakage, using equation 4.1. We indicated that we wanted R to print the result as well as storing it by enclosing the statement in parentheses. R calculates the result and give us the answer after an index number (here, `[1]`). We will also use the square brackets device `[2:1]` to reverse the order of printing some of the interval estimates to make them easier to read.

Consider the following values for a pathway that is subject to one level of screening and one of inspection.

```
> v = 10000
> i = 3000
> b = 30
> n_i = 100
> y_i = 5
```

```
> n_r = 300
> y_r = 5
```

The estimated leakage count is

```
> (l.hat = i * (y_i / n_i) + (v - i) * (y_r / n_r))
```

```
[1] 266.6667
```

The standard error of this quantity is computed by

```
> p_i = (y_i + 1) / (n_i + 2)
> p_r = (y_r + 1) / (n_r + 2)
> (s_l = sqrt(i^2 * p_i * (1 - p_i) / (n_i + 2) +
+             (v - i)^2 * p_r * (1 - p_r) / (n_r + 2)))
```

```
[1] 89.69113
```

The 95% confidence interval estimate for the leakage count is

```
> (l.int = i * p_i + (v - i) * p_r + c(-1,1) * 1.96 * s_l)
```

```
[1] 139.7488 491.3380
```

The estimated approach count is

```
> (a.hat = b + l.hat)
```

```
[1] 296.6667
```

and the interval estimate of the approach count is

```
> (a.int = b + l.int)
```

```
[1] 169.7488 521.3380
```

Then BIC is

```
> (BIC = (v - a.hat) / v)
```

```
[1] 0.9703333
```

and the interval estimate of BIC is

```
> (BIC.int = (v - a.int) / v)[2:1]
```

```
[1] 0.9478662 0.9830251
```

To get PIC we now compute the leakage from all intervention

```
> (L.hat = l.hat - y_r - y_i)
```

```
[1] 256.6667
```

and its interval estimate

```
> (L.int = l.int - y_r - y_i)
```

```
[1] 129.7488 481.3380
```

Then PIC is

```
> (PIC.hat = (v - L.hat) / v)
```

```
[1] 0.9743333
```

and its interval estimate is

```
> (PIC.int = (v - L.int) / v)[2:1]
```

```
[1] 0.9518662 0.9870251
```

The NCE of the inspection, expressed as a percentage, is

```
> (NCE.insp.hat = b / a.hat) * 100
```

```
[1] 10.11236
```

with interval estimate

```
> (NCE.insp.hat = b / a.int)[2:1] * 100
```

```
[1]  5.754424 17.673170
```

The NCE of the screening, expressed as a percentage, is

```
> (NCE.scr.hat = (b + i * (y_i / n_i))/ a.hat) * 100
```

```
[1] 60.67416
```

with interval estimate presented in Appendix B. Finally, the HR of the screening is computed using the estimated inspection leakage count

```
> (l_i = i * (y_i / n_i))
```

```
[1] 150
```

and its standard error

```
> (s_l_i = i * sqrt((y_i + 2) * (n_i - y_i + 2) / ((n_i + 4)^2 * n_i)))
```

```
[1] 75.16624
```

as follows, to obtain an interval estimate

```
> (l_i_int = i * (y_i + 2) / (n_i + 4) + c(-1, 1) * 1.96 * s_l_i)
```

```
[1]  54.59725 349.24890
```

Then, expressed as a percentage, the HR is

```
> (HR.hat = ((b + l_i) / i)) * 100
```

```
[1] 6
```

and its interval estimate is

```
> (HR.int = ((b + l_i_int) / i)) * 100
```

```
[1]  2.819908 12.641630
```

4.3 Aggregation and Dis-aggregation

4.3.1 Across Pathways

The measures above can be computed across pathways by simply summing the relevant quantities for each pathway, and then using the formulas above. Thus, for example, given two pathways A and B, the intervention leakage count across both pathways can be computed from $i = i_A + i_B$, $n = n_A + n_B$, and $y = y_A + y_B$. Treating the two pathways as strata, and formally summing the weighted means and weighted variances is also possible, but the algorithms become complicated and the benefit is uncertain.

4.3.2 Within Pathways

In theory, the performance indicators listed will scale easily to any sub-pathway level, as long as the data are available divided into sub-pathway statistics. However, determining the volume of sub-pathways can be a tricky proposition for both the international passengers and mail pathways. This difficulty arises from the fact that passenger numbers in each intervention method are determined by collecting and counting IPCs, and counting the IPCs by sub-pathway would be too time-consuming. A similar challenge faces the mail pathway. This problem is documented in **?**. The same report provides the proposed solution, which involves estimating the sub-pathway volumes using the intervention counts from the leakage survey and a statistical technique called *raking*. The calculation of appropriate interval estimates under these circumstances is under examination by ACERA.

Alternative Instruments for Estimating Sub-pathway Volume

Presently the leakage survey is used in two ways: to estimate the effectiveness of the intervention, and also to estimate the volume of different cohorts of units that are subject to different types of intervention (see **?**). The latter can then be used to estimate approach rates for combinations of cohorts and intervention types to develop profiles. This is an efficient use of the leakage survey data. If more data were needed for estimating passenger cohort intervention rates, then it would also be possible to perform a snapshot or sustained sample survey of IPCs. Counting a sample of a reasonable size within each channel would provide useful information; counting all the cards, especially in well-traveled regions, would be unnecessary. It would be useful to undertake a study to determine what is the effect on the statistical quality of the indicators of moving to a sample-based estimate of passenger card counts as a basis for estimating passenger volumes for pathways and sub-pathways.

4.4 Rates and Counts

Both BIC and PIC express the number of non-compliant passengers as a proportion of pathway volume. Hence if an estimate of the total number of non-compliant passengers passing through border control is required then it can be determined using pathway volume. For example, consider a pathway with a PIC of 0.99 and a volume of 100,000. For this pathway, 1000 non-compliant passengers are getting through. For another pathway with a PIC of 0.95, but with a volume of 10,000 only 500 non-compliant passengers are getting through. This gives an estimate of post-border incursions into Australia and provides useful context.

4.5 Handling Small Cohorts

The original intention of the leakage survey was to enable an estimate of the leakage rate through various types of intervention, for example, inspection, screening, and so on. More importantly,

some of the passengers are released after screening of the IPC, and the leakage survey is the only way of obtaining information about these passengers. However, a further use has been found. **?** and **?** recommended using the data that are collected during the leakage survey to estimate the number of passengers of each passenger cohort processed by each intervention method. Therefore both of the leakage survey outcomes, namely n and y, are useful for estimating performance indicators in the passengers and mail pathways.

There is a possibility that y could be zero when n is small, even if leakage is present, and consequently there would appear to be no leakage. Presently, cohorts that are sufficiently small are merged or clustered into groups, which is a reasonable although time-consuming and arbitrary strategy. Clustering is an example of making a 'bias–variance' trade-off; when we cluster poorly represented cohorts, we accept the possibility of considerable bias in return for a reduction in variance. For example, we could assume that all passenger cohorts present the same leakage rate, and estimate one leakage rate for each intervention method.

An alternative is to reason as follows. The leakage rate is estimated as a binomial random variable, so we advocate using the inverse-cumulative distribution function for the beta distribution to obtain estimates and confidence intervals of the estimated proportion. The parameters of the beta distribution could follow the Clopper–Pearson approach, namely $\text{BetaInv}(0.025, y, n - y + 1)$ for the lower limit of the interval and $\text{BetaInv}(0.975, y + 1, n - y)$ for the upper limit of the interval, or the Jeffreys Prior approach, which is $\text{BetaInv}(q, y + 0.5, n - y + 0.5)$ for any quantile q. The choice between these two approaches could be made based on receiver operating characteristic (ROC) curves or leakage curves. This strategy would tend to increase the targeting upon small cohorts about which little is known. In doing so, this targeting would follow the principle that ignorance is also a source of risk, as proposed in earlier ACERA projects (for example **?**).

For cohorts for which no leakage surveys were taken at all (*ie* $n = 0$), and for which no auxiliary information is available, we advocate substituting the average leakage across all cohorts. On average such cohorts would be very small, so we would argue that operational simplicity should overshadow statistical accuracy. Alternatively, small cohorts might also be assumed to be automatically targeted, in order to gather information preferentially about them. If auxiliary information is available then it might be used to justify a higher risk rating.

Another alternative that would allow seamless handling of the poorly-represented cohorts would be to use an empirical Bayes strategy, as outlined in **?**. It would be useful to assess the impacts of each of these kinds of strategies upon the statistical and operational qualities of the performance indicators.

4.6 Other Ways to Measure Leakage

DAFF has sometimes considered a different leakage estimation strategy to that presented above. We discuss the alternative in this section.

Recall that the leakage surveys performed for passengers and mail by DAFF involve the random, manual inspection of a unit (mail article or passenger) that has been cleared by biosecurity intervention methods. In the passenger pathway the manual inspection is of one unopened bag, randomly selected. A similar approach is presently used by NZ MPI in airports.

The leakage survey faces some non-statistical issues. First, the leakage unit selection is designed to be as random as possible within the facility, but it is difficult to guarantee. There are apparent disincentives for truly random selection; for example an officer may avoid a passenger that has many bags in favour of a passenger that has just one. Second, there are concerns about collegiality; that is, inspectors may be reluctant to report on a colleague's performance. Third, in the passengers pathway, the leakage survey is a burden on the passengers that have already been cleared. Obviously this burden is less of a concern in the mail pathway.

The alternative to the *leakage* survey can be termed an *approach* survey, which typically involves a random inspection of a sample of all units *before intervention*. The seizure rate is

known from the inspection history. Then either the leakage rate or the approach rate needs to be measured in order to estimate the other rate.

The approach survey is statistically less efficient than the leakage survey for estimating leakage, so a larger number of random inspections would need to be made in order to achieve estimates of the same statistical precision. Also, it is possible that NCE estimates computed this way could be greater than 1. Finally, the approach survey gives no insight to what is happening in the intervention sub-processes, e.g. it cannot provide NCE and Hit Rate statistics that are specific to screening by dogs, etc. However, the approach does enjoy some other benefits. Achieving a random sample is arguably easier, because it can be flagged before any biosecurity intervention. Also, in this setting there are no disincentives for detection, as there are no concerns about collegiality.

The choice between the two approaches depends on the following points: the leakage survey used for the passengers and mail pathways is more efficient in its use of inspection resources, but the approach survey avoids the reporting bias that complicates interpretation of the leakage estimates that arise from the leakage rate survey. Neither approach is free of sampling bias, in which a more complicated inspection might be passed over for a simpler one.

Note that the passenger information collected during the leakage survey is also used for the performance indicators. If the approach survey were to be used instead, alternative means to gather the passenger information would be needed, for example, analysis of the IPCs.

4.7 Recommendation

We recommend that DAFF adopt the performance indicators described in this report for the international passengers and mail pathways, and begin to try to use them in other pathways as deemed useful by pathway managers.

Selection and adoption of performance indicators is partially a matter of organizational culture. The overarching goal is to assess the organization's performance. The assessment should cover simple principles such as efficiency and effectiveness of intervention. However, different indicators can be used to represent these simple principles. For example, this report recommends hit rate as a measure of screening efficiency. Alternatives to the hit rate include the odds ratio and the false positive rate. Each is interpreted differently and has different statistical properties. We recommend that DAFF review the choice of indicators and data collection procedures within a year.

5

Profiling

5.1 Introduction

Profiling is the division of a pathway into easily defined sub-pathways (cohorts) that have different levels of risk (BIC). For example, DAFF might have historical data that suggest that mail articles arriving from a specific country, or passengers on a specific flight, are more likely to be non-compliant than mail articles arriving from other countries or passengers on other flights. Profiling allows the inspectorate to focus its intervention resources on the highest risk (lowest compliance) cohorts under its purview.

Profiling for the international passengers and mail pathways is a two-step process. First, the profiling is used to identify which cohorts of the pathway have the highest proportions of non-compliance. Second, a decision must be made as to the intervention method to be used for each cohort. Usually there are three choices: manual inspection, detector dogs, and X-ray units.

In this chapter we briefly review the existing strategies for constructing and using national profiles, and then outline how these strategies should change to reflect the introduction of the proposed performance indicators.

5.2 Current Strategy

5.2.1 Identification of Risky Cohorts

The current strategy for profiling follows that described in **?** and **?**. Briefly, units are divided into cohorts (here, citizenship and flight) and the cohorts are prioritized by their estimated approach rates. We recommend that this profiling approach be retained. It will produce profiles that are identical to those produced using, for example, BIC, which can also be used.

The process of obtaining the estimated approach rate is complicated by the fact that units undergo different types of intervention, namely assess and release, X-ray, detector dogs, and manual inspection, to detect non-compliance. These different types of intervention each have a different probability of detecting non-compliance. In order to obtain a reasonable estimate of the approach rate for each cohort, we need to know how many of each cohort is processed by each intervention type. However, obtaining this information would require counting the number of units for each cohort within each intervention type, which is time consuming. Therefore estimates of the number of each cohort processed in each intervention type are calculated using the leakage survey and a statistical technique called *raking*, also known as iterative proportional fitting (see **?**).

In addition to the necessity for statistical modelling, construction of the profiles requires matching of passenger citizenship between Customs and Border Protection records and MAPS, the DAFF database that is used to record intervention results for passengers and mail. The overlap between most citizenships is clear, but some do not match. Decisions must be made on

how to map such categories, which in some cases may entail some clustering. Also there are many small volume cohorts for which the subsequent calculations may be problematic. Strategic clustering to combine smaller cohorts is then needed. It would be useful to undertake a study on when to cluster, how to cluster, and how to assess the effect upon the profiles of the clustering.

5.2.2 Screening Choice

The screening choice for high-risk cohorts is performed by assessing the nature of the non-compliance seized from each cohort. We assume that the detection of specific types of non-compliance is either more likely with dog screening, X-ray screening, both, or neither. For example, if the non-compliance seized from a cohort is predominantly compatible with detection by dogs, then the cohort should be profiled for screening with dogs.

This approach is a useful way to advance a profiling choice, however, it suffers from the same disadvantage that identifying risky cohorts does: the need to avoid circularity of detection and self-fulfilling profiles. In order to be sure that the best possible profiles are used, the non-compliance that is detected should be weighted by the number of units processed by each method. These quantities are available from the analysis used to identify which were the riskiest cohorts.

A recent focus on screening efficiency (as in, cost per passenger screened) suggested that dog screening was substantially more efficient than X-ray machines, in that the cost per passenger of screening by dogs was substantially less than the cost per passenger of screening by X-ray. It may well be that this observation overrides other considerations.

5.3 Modifications On New Performance Indicators

Targeting those categories that have the highest approach rate (or lowest compliance rate) is an obvious way to intercept as many non-compliant items as possible. However, if the effectiveness of intervention is significantly different for some categories, then the expected seizure rate (ESR = $(1 - \text{BIC}) \times \text{NCE}$) may be a more appropriate criterion to base the profiles on. One other choice for either method is whether to base the category rankings on the estimates of the approach / seizure rates or on upper estimates, for example upper confidence levels. A comparison of BIC and ESR (not shown here) suggests that the differences are not important.

6

Reporting

We now provide examples of the new performance indicators, computed using DAFF and Customs and Border Protection data from August 2010 to July 2011 inclusive. Note that the data holdings for that period were not sufficiently detailed to compute Hit Rate as formally defined above, so we have computed a direct ratio of the number of non-compliant units identified over the number of units screened.

The reader should note that, at present, we assume that the only statistical source of uncertainty in the estimates is from the leakage survey: the leakage survey is a sample of the passengers, hence the true leakage is unknown. The uncertainty that this creates is propagated through the other estimates, hence BIC, PIC, and NCE are all accompanied by confidence intervals. These intervals will shrink as more data become available. A full accounting of the uncertainty would include recognition that the cohort volumes (such as citizenship) are not known by intervention type, and have been estimated following the methods laid out in **?**.

Here we present summaries by intervention method, declaration, and port, and also by citizenship. We also examined the statistics by month but found them too variable to yield useful information. This point could be revisited when more data become available.

6.1 Intervention Method and Port

Figure 6.1 shows the BIC for the year, computed by region, declaration status, and intervention method. The interval widths provide relative information on how reliable the estimates are: smaller intervals correspond to greater reliability. Note that the estimate can be at the border of the confidence interval, for example when the estimate of leakage is 0. The figure shows that, with the exception of the manual inspection intervention method, there is not a substantial difference between the approach rates of the declarant and non-declarant passengers. The differences between the *Assess and Release* and the other three intervention methods within both the declarant and non-declarant columns are a measure of how well the passengers are profiled. The lower BIC in the *Non-declarant Manual* panel suggests that the assessment using profiles is having some effect.

Figure 6.2 shows the PIC for the year, computed by region, declaration status, and intervention method. Notably low areas are Declarant Assess and Release and Declarant dogs screening for Port 6, and Non-declarant X-ray and Non-declarant Manual Inspection for Ports 5 and 2 respectively. The confidence intervals are wide, however, so the patterns may be due to random fluctuations in the leakage survey.

Figure 6.3 provides the NCE for the year, computed by region, declaration status, and intervention method. Note that the effectiveness for Assess and Release is 0. The width of many of the confidence intervals show that the sample sizes are quite small for a number of the intervention methods. Where the estimates are more precise, the estimates are also low; for example the effectiveness for Declarant Manual Inspection and Declarant X-ray Inspection

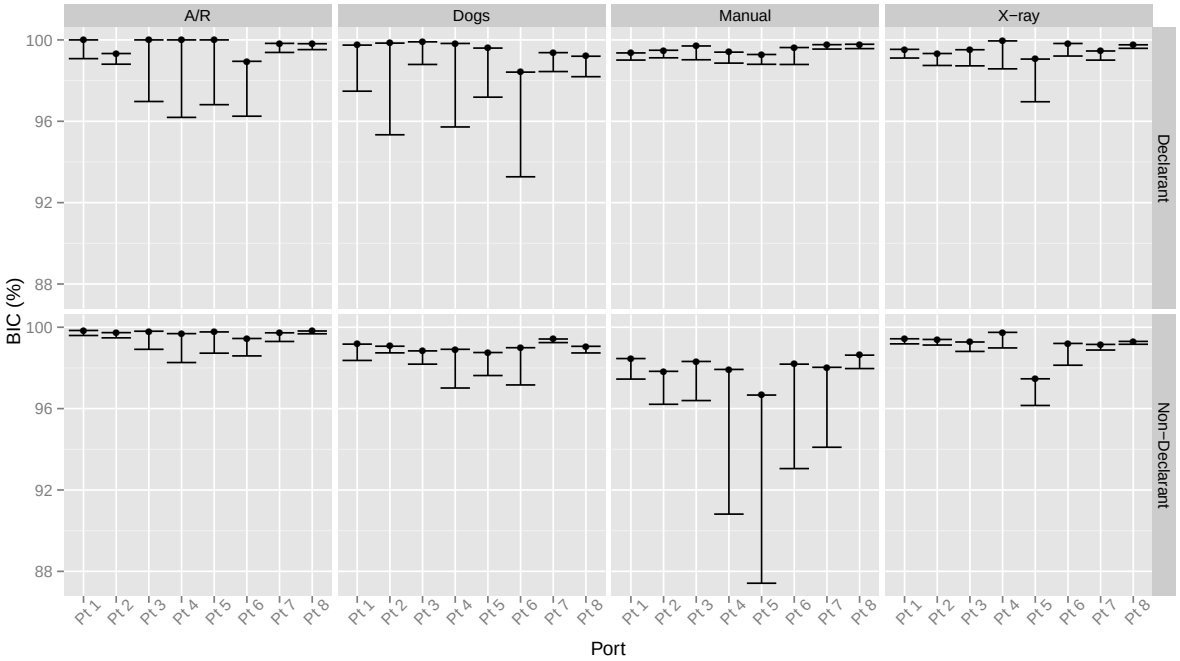

Figure 6.1: Before-Intervention Compliance (BIC) categorized by declaration status (rows), intervention method (columns), and airport (x-axis). A/R is Assess and Release, K9 refers to dogs, Manual means unpack and inspect, and Xray is as labeled.

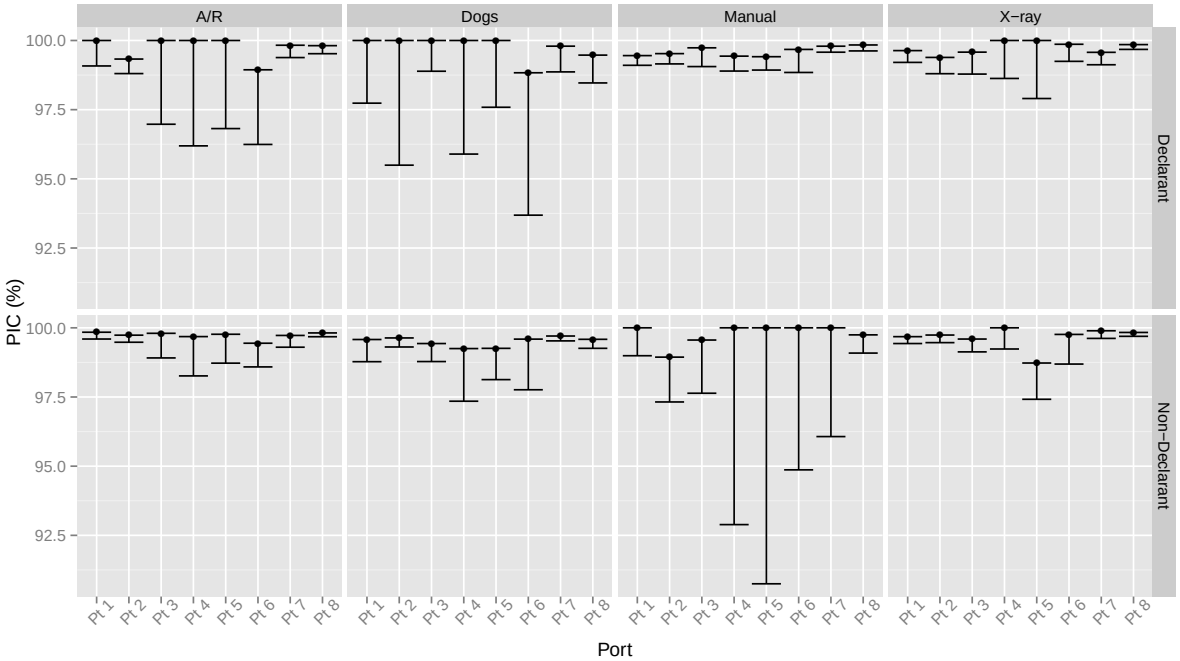

Figure 6.2: Post-Intervention Compliance (PIC) categorized by declaration status (rows), intervention method (columns), and airport (x-axis). See caption of Figure 6.1 for key.

are quite low, especially compared with the effectiveness for non-declarants in each case. More data are needed, but, coupled with Figure 6.1, this result suggests that the *Declarant Manual* intervention method is reasonably clean of non-compliance, but the non-compliance that is there is not getting detected effectively.

Figure 6.4 gives the inspection-based Hit Rate for the year, computed by region, declaration

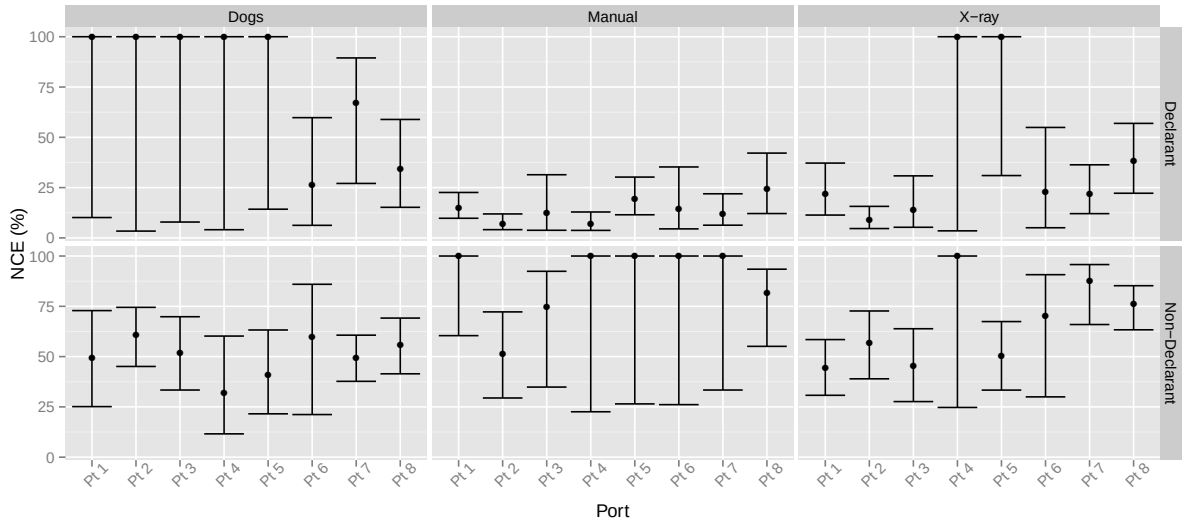

Figure 6.3: Effectiveness (NCE) categorized by declaration status (rows), intervention method (columns), and airport (x-axis). See caption of Figure 6.1 for key.

status, and intervention method. The graph has limitations because the data do not distinguish between interceptions from referrals to manual inspection from screening by dogs and X-ray, and by profiling from the RAO. Also, under these data we do not know whether a manual inspection actually searched for non-compliance or simply examined declared goods. The motivation for the manual inspection is unknown. Taking account of these caveats, this figure shows that Hit Rate is highest in Non-Declarant Manual inspections from Port 5. The Declarant Manual and X-ray Hit Rates are low, with the exception of X-ray in Port 5, which is at least five times higher than the other regions.

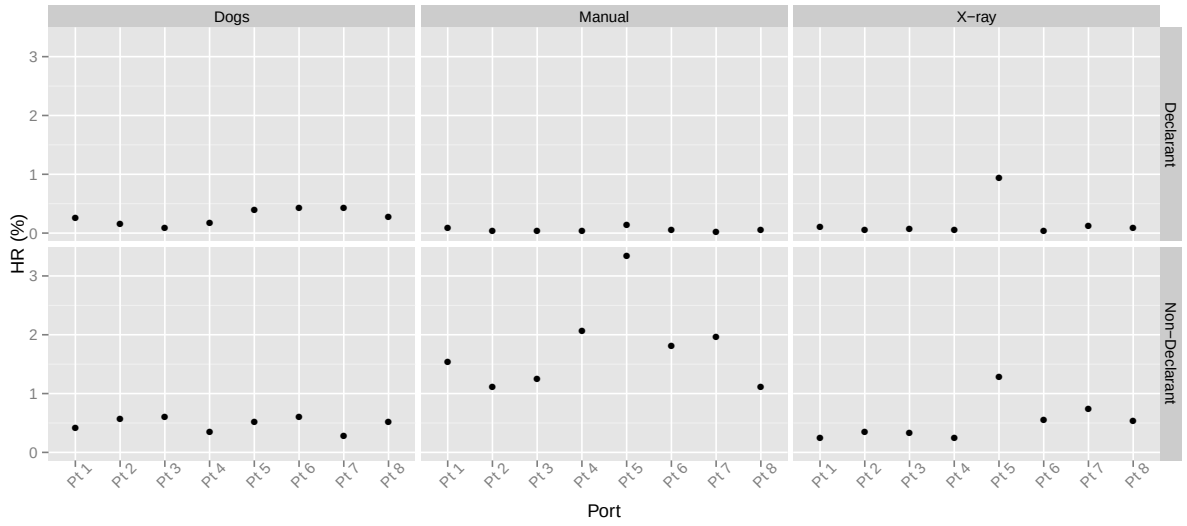

Figure 6.4: Hit Rate categorized by declaration status (rows), intervention method (columns), and airport (x-axis). Note that the Hit Rate for Assess & Release cannot be computed.

This collection of figures could be further augmented with similar graphs of estimated leakage count and approach count, but we do not produce them for this public report.

6.2 Process

Figure 6.5 provides a collection of all four indices by process. The BIC panel (top left) shows that assessment using profiles is having an effect: namely, the BIC is lower for the Non-Declarant dogs, X-ray and Manual Inspection intervention methods than the Assess and Release intervention method. A smaller degree of success exists for assessment using profiles for declarants. The PIC panel shows that the PICs are about the same for all the intervention methods. Effectiveness (NCE) is high for Declarant screening by dogs and the three non-declarant inspection-based intervention methods. A similar pattern is found for the Hit Rate, with non-declarant manual Hit Rate around three times higher than any other, although recall that the Hit Rate statistics presented here are based on inappropriate data, and are included to provide an indication of the kinds of information that will come available.

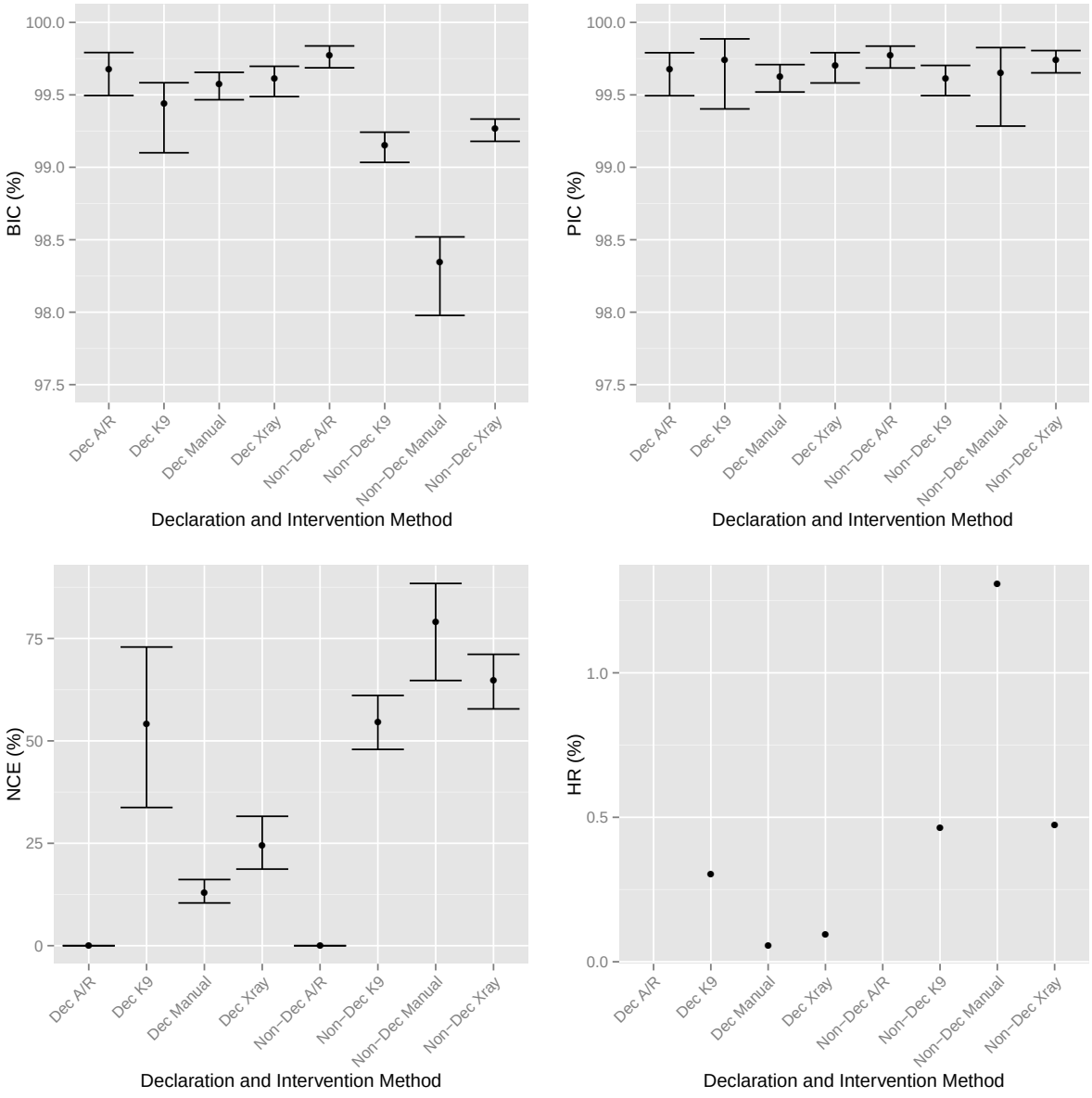

Figure 6.5: BIC, PIC, NCE, and Hit Rates by intervention method, including confidence intervals.

6.3 Citizenship

Figure 6.6 provides a collection of all four indices by citizenship, for the ten cohorts with the highest numbers of passengers processed. The BIC is low for M3 and P6 passengers, and high for N7, F1, and F4 passengers. PIC is highest for F5 and N7 passengers, and lowest for M3 and P6 passengers. The reason for the high PIC for F5 passengers is the high NCE. Effectiveness is also good for P6 and B9 passengers. The inspection-based Hit Rate is low in all cases, but relatively higher for M3, F5, and P6 passengers.

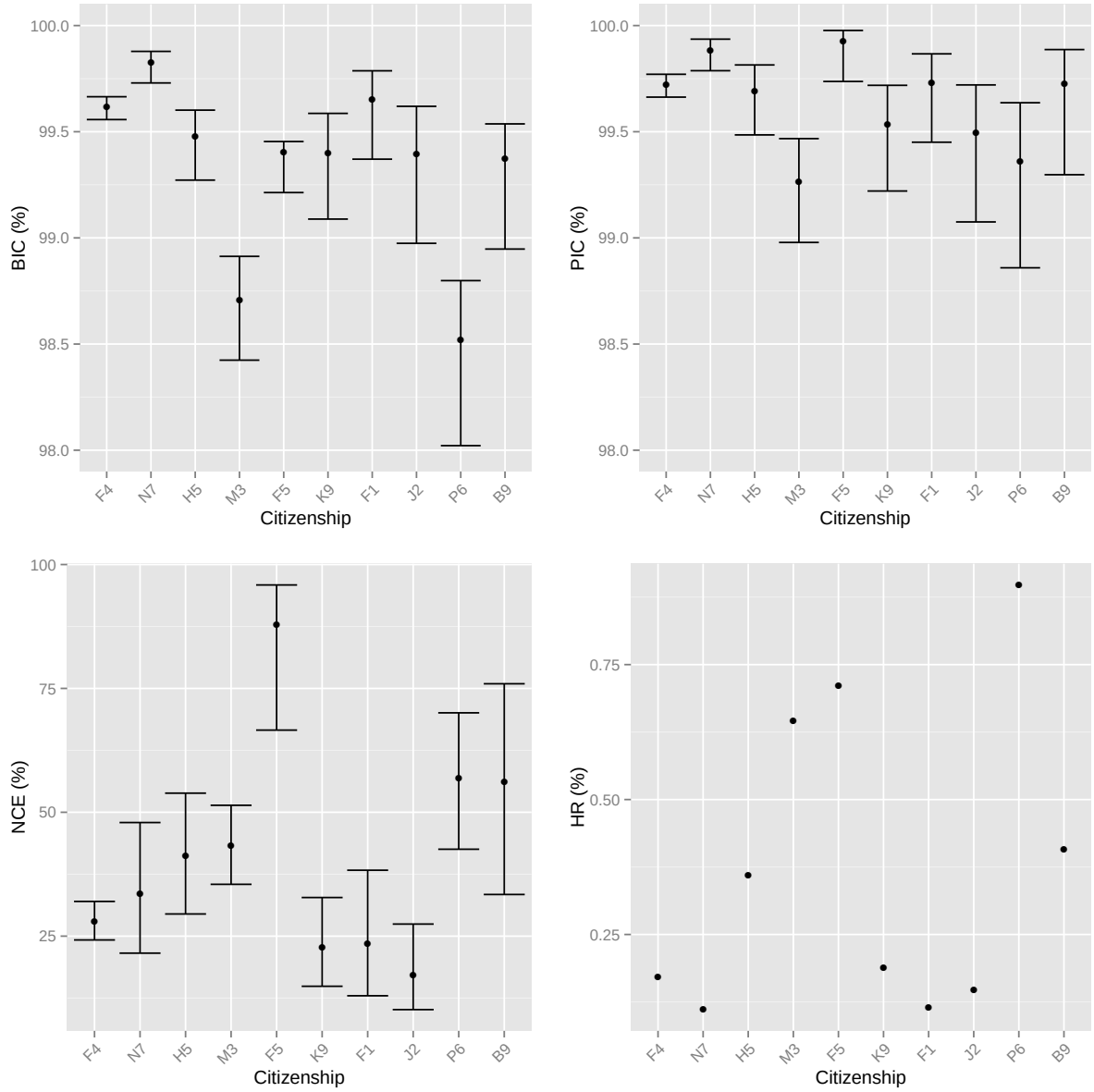

Figure 6.6: BIC, PIC, NCE, and Hit Rates by citizenship, including confidence intervals, for the top ten citizenships by passenger count, sorted by decreasing passenger count.

Figure 6.7 provides a summary of BIC by citizenship for all those citizenships that had 100 or more passengers included in the leakage survey, presented in decreasing order of the number of passengers processed. The N2 passengers are shown to have an unusually low BIC.

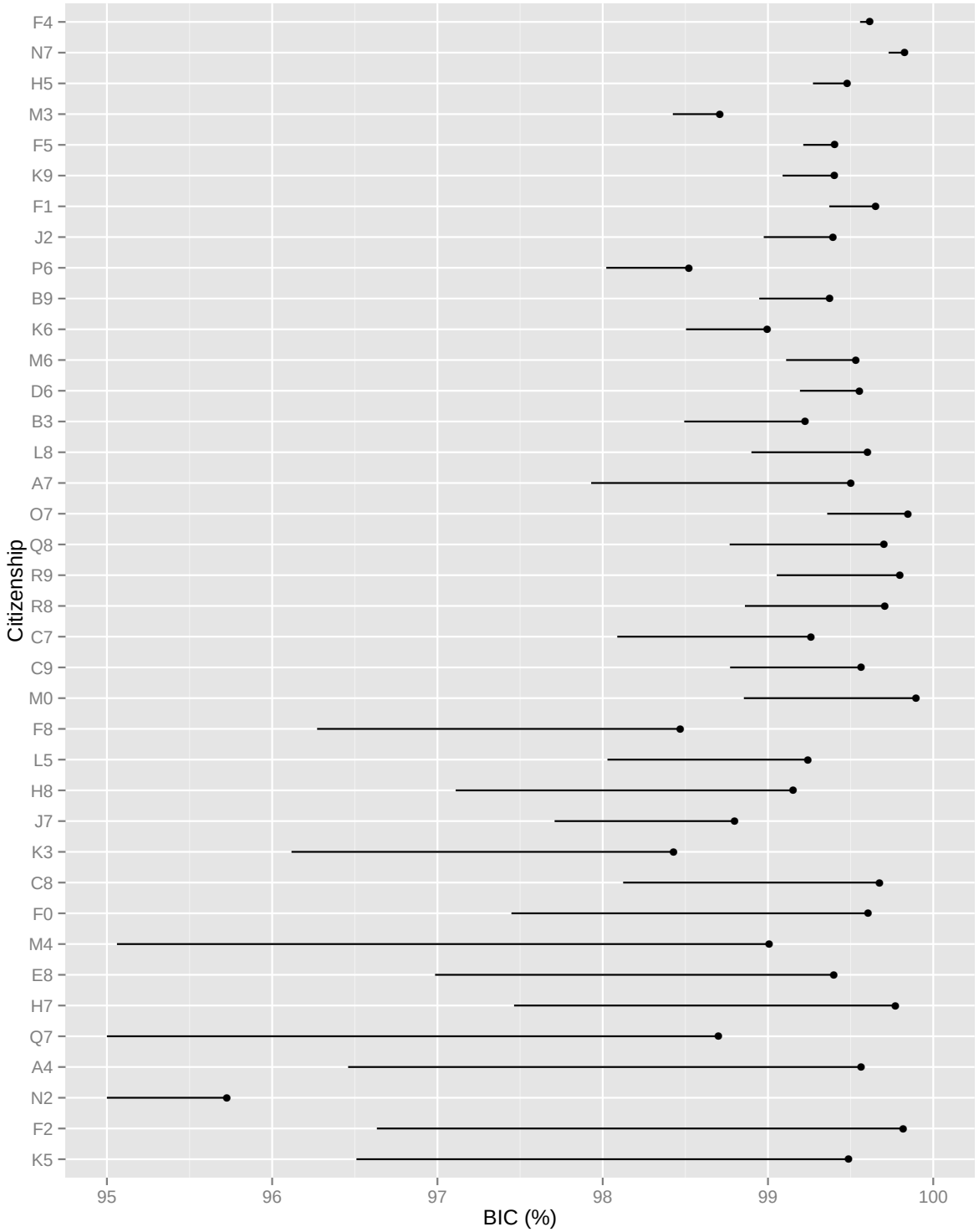

Figure 6.7: BIC by citizenship, including lower confidence intervals, for the citizenships with more than 100 passengers in leakage survey, sorted by decreasing passenger count. Some intervals are truncated.

Figure 6.8 provides a summary of BIC by citizenship for all those citizenships that corresponded to the lowest BICs, presented in increasing order of BIC. One way to establish a cutoff

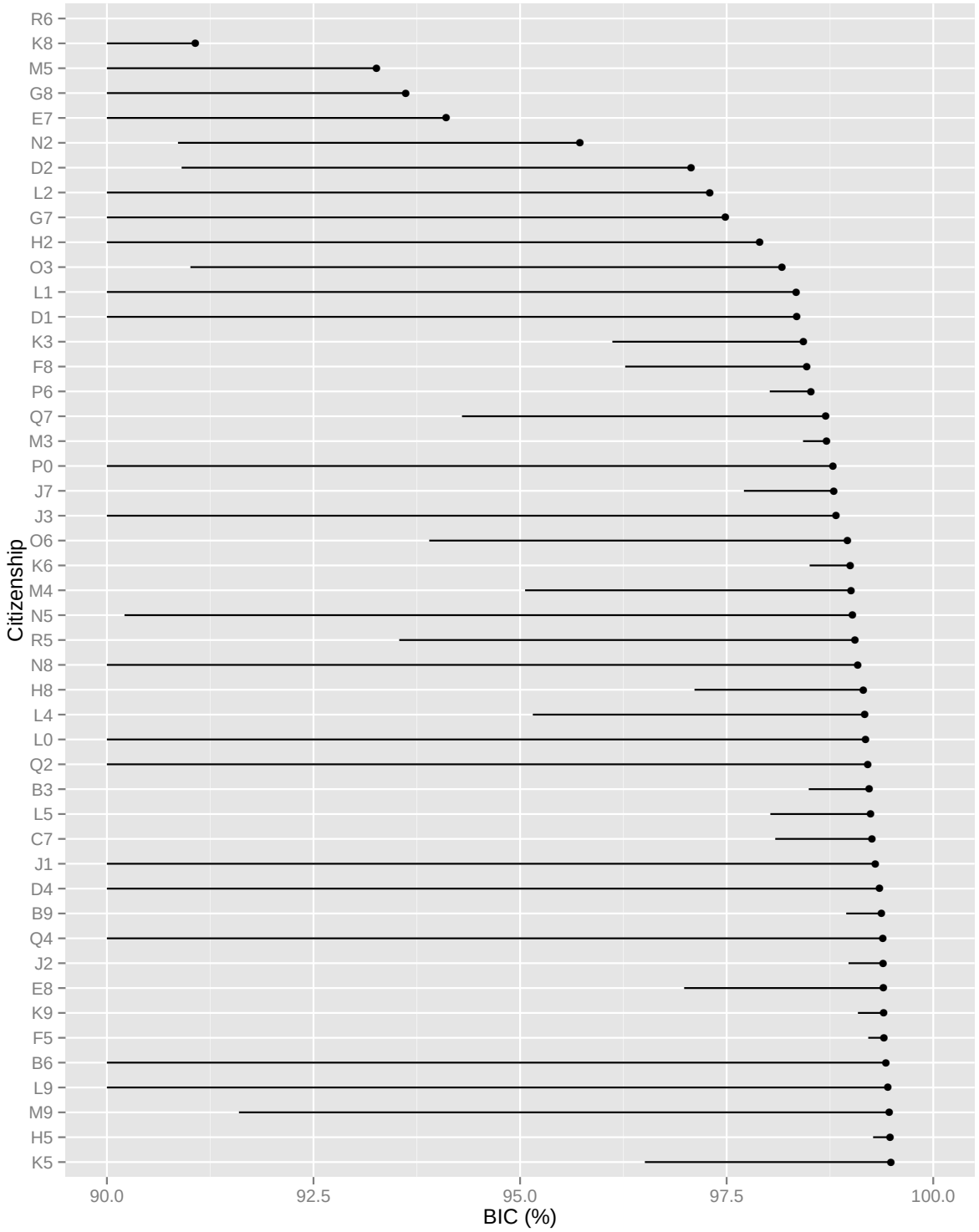

Figure 6.8: BIC by citizenship, including lower confidence interval, for the citizenships with the lowest BIC, in increasing order. Some intervals are truncated.

for determining which cohorts to intervene with would be to draw a vertical line and intervene with all citizenships that are to the left of the line. An alternative would be to apply the same approach but to represent each cohort by the lowest end of the confidence interval of the BIC.

6.4 Leakage Survey

We conclude the reporting example with some summary statistics about the leakage survey. The purpose of these statistics is to provide some insight as to whether the leakage survey is representative of the passenger cohorts, and how much data is available for drawing conclusions about inspection effectiveness at different levels of aggregation. These statistics are important for assessing how well the leakage survey is carried out. DAFF should investigate how to assess and report the representativeness of the leakage survey.

Table 6.1 provides the number of passengers included in the leakage survey for each intervention method and region, and Table 6.2 provides the percentage of passengers included in the leakage survey for each intervention method and region. Similarly, Table 6.3 provides the number of passengers included in the leakage survey for each citizenship and region, and Table 6.4 provides the percentage of passengers included in the leakage survey for each citizenship and region. Taken together, these tables show no evidence of non-representativeness in the sample across either intervention method or citizenship. The proportions vary somewhat, especially in the smaller regions, but not to the extent that the variation suggests any concerns. It is possible that a formal statistical summary could be constructed to estimate the representativeness of the leakage survey, for the purposes of assessing the system performance.

Table 6.1: Number of passengers in the leakage survey, by region and intervention method.

Intervention	Pt 1	Pt 2	Pt 3	Pt 4	Pt 5	Pt 6	Pt 7	Pt 8
Declarant								
A/R	399	1644	120	95	114	190	1169	2144
Dogs	161	80	330	88	151	86	489	569
Manual	2762	2327	767	1432	1728	624	3413	3096
X-Ray	1650	1312	723	267	174	738	1647	4052
Non-Declarant								
A/R	2530	3025	511	319	435	724	1462	6522
Dogs	715	2468	1077	271	547	247	5789	2657
Manual	364	380	235	50	38	70	92	790
X-Ray	3480	2699	1515	480	559	424	1892	5996

Table 6.2: The percentage of arriving passengers included in the leakage survey, summarized by region and intervention method.

Intervention	Pt 1	Pt 2	Pt 3	Pt 4	Pt 5	Pt 6	Pt 7	Pt 8
Declarant								
A/R	0.49	0.33	1.17	1.05	0.94	0.65	0.51	0.34
Dogs	0.41	0.27	1.25	1.02	0.81	0.34	0.41	0.27
Manual	0.63	0.45	1.57	1.32	1.24	0.82	0.66	0.44
X-Ray	0.63	0.44	1.57	1.33	0.95	0.80	0.73	0.43
Non-Declarant								
A/R	0.60	0.40	1.45	1.24	1.16	0.79	0.63	0.42
Dogs	0.68	0.34	1.99	1.62	1.41	0.92	0.83	0.37
Manual	0.91	0.63	3.78	2.30	0.72	1.04	0.50	0.60
X-Ray	0.97	0.58	2.37	1.89	1.58	1.09	0.86	0.49

Table 6.3: Number of passengers in the leakage survey, by region and citizenship, for top 15 citizenships in terms of count of passengers processed.

Citizenship	Pt 1	Pt 2	Pt 3	Pt 4	Pt 5	Pt 6	Pt 7	Pt 8
B3	74	90	50	31	18	3	149	433
B9	122	167	126	6	51	3	243	740
D6	130	103	34	50	30	2	200	473
F1	726	703	29	56	218	320	191	387
F4	5632	6025	1913	2014	1919	1036	7144	11218
F5	159	420	151	30	21	5	497	1463
H5	903	667	313	118	221	48	911	1390
J2	626	446	59	60	131	16	251	403
K6	182	186	88	84	57	1	263	406
K9	252	119	1236	14	43	731	187	441
L8	142	161	16	16	24	3	133	244
M3	414	1216	138	32	356	30	649	1944
M6	564	239	7	69	46	9	77	349
N7	426	1435	397	71	207	859	2871	2407
P6	269	555	26	25	146	8	241	457

Table 6.4: Percentage of arriving passengers in the leakage survey, by region and citizenship, for top 15 citizenships in terms of count of passengers processed.

Citizenship	Pt 1	Pt 2	Pt 3	Pt 4	Pt 5	Pt 6	Pt 7	Pt 8
B3	0.80	0.39	2.20	1.75	1.43	0.32	0.75	0.44
B9	0.80	0.47	2.14	1.27	1.24	0.59	0.68	0.47
D6	0.86	0.35	2.23	1.66	1.72	0.43	0.89	0.50
F1	0.73	0.50	2.05	1.49	1.35	0.94	0.78	0.49
F4	0.65	0.41	1.79	1.33	1.18	0.84	0.68	0.42
F5	1.01	0.47	2.86	2.07	1.65	0.84	0.94	0.48
H5	0.63	0.37	1.91	1.44	1.23	0.85	0.71	0.40
J2	0.79	0.56	1.96	1.66	1.48	0.97	0.93	0.54
K6	0.94	0.40	2.76	2.26	1.42	0.25	1.03	0.45
K9	0.62	0.39	1.53	1.14	1.01	0.76	0.61	0.37
L8	0.70	0.39	1.74	1.22	1.08	0.30	0.74	0.42
M3	1.03	0.49	2.47	1.77	1.68	0.72	0.81	0.48
M6	0.93	0.53	2.33	1.80	1.53	0.77	0.90	0.54
N7	0.64	0.40	1.74	1.35	1.19	0.87	0.70	0.44
P6	1.16	0.47	3.18	1.96	1.47	0.44	0.84	0.49

7

Conclusion and Recommendations

7.1 Overview

This report provided examples of applications of the recently developed performance indicators in the international passengers pathway.

7.1.1 Prior Work

DAFF has adopted a risk-based approach to managing the biosecurity risk of various pathways, including international passengers and mail. Following the Australian National Audit Office's 2001 report, inspection effectiveness was used as the primary indicator of inspectorate performance. However, with the implementation of a risk-based approach to management, a richer suite of indicators is required. ACERA Project 1001i recommended *post-intervention compliance* (PIC) of the pathway as a performance indicator. The goal of the current project is to demonstrate the use of the indicators and to assess possibilities for broadening their use.

7.1.2 This Report

This project focused on broadening the scope of the indicators, implementing them for the international passengers pathway, and assessing the effect on prioritization of passenger cohorts for further intervention. The four indicators described in this report are:

1. Before Intervention Compliance (BIC),
2. Post Intervention Compliance (PIC),
3. Non-Compliance Effectiveness, and
4. Hit Rate.

This report provided examples of computing and reporting these indicators by region, intervention method, and declaration status. The report also provided examples of monthly summaries, which may be too variable to yield useful information.

7.2 Recommendations

The recommendations arising from this report are as follows.

1. The proposed performance indicators (BIC, PIC, NCE, HR) should be used to assess how appropriately the inspectorate performs as well as how well it performs, with PIC as the key indicator (p 23).

2. Profiles for international passengers and mail articles should still be based on the approach rate (p 24).

3. Performance indicators should be reported with confidence intervals wherever possible, so that the manager can accurately assess the quality of the available information (p 14).

4. The nominal coverage of the confidence intervals should be no less than 90% (p 14).

5. DAFF should determine what would be the effect upon the statistical qualities of the performance indicators of using a sampling approach to counting Incoming Passenger Cards (IPCs) instead of counting all of them (p 21).

6. DAFF should undertake a further study to determine when and how to cluster small cohorts, and what is the effect upon profiling of that clustering, and what other options—for example, empirical Bayes estimates—might be available for handling small cohorts (p 22).

7. DAFF should consider whether the cutoff for targeting high-risk cohorts should be the mean approach rate or some higher confidence interval (or, BIC or lower interval). The mean is the best indicator of compliance, but the higher confidence interval acknowledges that ignorance is a source of risk (p 32).

8. Leakage surveys need to be representative in order to reduce uncertainty about the compliance of individual cohorts. DAFF should investigate how to assess and report the representativeness of the leakage survey (p 33).

9. DAFF should review the choice of performance indicators and the data collection procedures within one year (p 23).

Appendix A

Glossary

A.1 Important Acronyms

APHIS	USDA Animal and Plant Health Inspection Service
BIC	Before-Intervention Compliance
DDU	Detector Dog Unit
HR	Hit Rate
HRAO	Hall RAO
IPC	Incoming Passenger Card
IQI	Increased Quarantine Intervention
NCE	Non-compliance Effectiveness
NZ MPI	New Zealand Ministry for Primary Industries
PIC	Post-Intervention Compliance
RAO	Risk Assessment Officer
USDA	United States Department of Agriculture

Appendix B

Confidence Intervals for NCE for Screening

A stream of passengers with an approach rate of a is subject to a two-stage intervention process, namely a screening followed by a possible inspection, depending on the outcome of the screening. First, all items are screened by a process with an effectiveness of NCE_s. A number of these items will be referred for inspection, which has effectiveness NCE_i, and the rest are released. The inspection detects b passengers that are non-compliant, but may miss some. The non-compliant goods are rectified. The leakage survey records whether the passenger is in the inspected or released stream. Hence the leakage survey provides four counts: y_r finds from the n_r passengers that were released, and y_i finds from the n_i passengers that were inspected.

As mentioned in Section 4.1.4, the problem with estimating NCE_s for such a situation is that we don't know how many items were detected by the screening process, only the number of items detected by the screening process *and* by the inspection. We have to augment the number detected by inspection with the number missed by inspection. Let the subscript i refer to statistics from the *inspected* pathway, and r refer to the *released* pathway. Then,

$$NCE_s = \frac{b + \hat{l}_i}{b + \hat{l}_i + \hat{l}_r} \tag{B.1}$$

This presents a challenge for interval estimation because the estimate of NCE_s is a ratio of two non-independent random variables. We present two approaches, both of which start by rearranging equation B.1 so that we can express NCE_s as the ratio of two independent random variables.

B.1 Direct Method

We acknowledge and appreciate the contribution of Rob Cannon, a reviewer, to the material contained in this section. When we express equation B.1 as

$$NCE_s = \frac{(b + \hat{l}_i)}{(b + \hat{l}_i) + \hat{l}_r} \tag{B.2}$$

we can see that it comprises two parts, say x and y, in the pattern $NCE_s = x/(x + y)$. Let $x = b + \hat{l}_i$ and $y = \hat{l}_r$. Immediately we can see that x and y are independent. Further, we will invoke the Central Limit Theorem and assume that they are normally distributed, that is,

$$x = b + \hat{l}_i \sim \mathcal{N}(b + \mu_{l_i}, \sigma^2_{l_i}). \tag{B.3}$$

and

$$y = \hat{l}_r \sim \mathcal{N}(\mu_{l_r}, \sigma^2_{l_r}) \tag{B.4}$$

Now use e in place of NCE_s for convenience. We can then rewrite the equality in equation B.2 as

$$\begin{aligned} \frac{x}{x+y} &= e \\ x &= e \times (x+y) \\ x - e \times x - e \times y &= 0 \\ x \times (1-e) - y \times e &= 0 \end{aligned} \tag{B.5}$$

The left-hand side of (B.5) is a random variable, and is normally distributed (conditional on e) because we assumed that x and y are both normal. That is,

$$x \times (1-e) - y \times e \sim \mathcal{N}(\mu_x \times (1-e) - \mu_y \times e, \sigma^2_x \times (1-e)^2 + \sigma^2_y \times e^2) \tag{B.6}$$

We now construct an interval for e as follows. Values of e that are unlikely are those that make the random variable $x \times (1-e) - y \times e$ far from zero. If we construct a 95% confidence interval for that random variable, and standardize it, then we can express that interval in terms of e. We then solve the interval limits for e. First, we construct the 95% confidence interval for the standardized random variable (NB: the mean is zero).

$$\frac{\bar{x} \times (1-e) - \bar{y} \times e - 0}{\sqrt{\sigma^2_x \times (1-e)^2 + \sigma^2_y \times e^2}} = \pm 1.96 \tag{B.7}$$

Squaring both sides and simplifying leads to

$$(\mu^2_x - 1.96^2\sigma^2_x)(1-e)^2 - 2e(1-e)\mu_x\mu_y + (\mu^2_y - 1.96^2\sigma^2_y)e^2 = 0 \tag{B.8}$$

Now divide both sides by e^2 to get the following quadratic in $r = \frac{1-e}{e}$:

$$(\mu^2_x - 1.96^2\sigma^2_x)\left(\frac{1-e}{e}\right)^2 - 2\mu_x\mu_y\left(\frac{1-e}{e}\right) + (\mu^2_y - 1.96^2\sigma^2_y) = 0 \tag{B.9}$$

Recall that the roots of the quadratic equation $Ar^2 + Br + C = 0$ are

$$r = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \tag{B.10}$$

Here, we have the following quantities

$$\begin{aligned} A &= \mu^2_x - 1.96^2\sigma^2_x \\ B &= -2\mu_x\mu_y \\ C &= \mu^2_y - 1.96^2\sigma^2_y \end{aligned} \tag{B.11}$$

so the roots are

$$\frac{2\mu_x\mu_y \pm \sqrt{4\mu^2_x\mu^2_y - 4(\mu^2_x - 1.96^2\sigma^2_x)(\mu^2_y - 1.96^2\sigma^2_y)}}{2(\mu^2_x - 1.96^2\sigma^2_x)} \tag{B.12}$$

which simplifies slightly to

$$\frac{\mu_x\mu_y \pm 1.96\sqrt{\mu_y^2\sigma_x^2 + \mu_x^2\sigma_y^2 - 1.96^2\sigma_x^2\sigma_y^2}}{\mu_x^2 - 1.96^2\sigma_x^2}. \tag{B.13}$$

The roots finally need to be back-transformed by $e = 1/(1+r)$. The back-transformed interval is biased, and it may be worth trying to estimate and correct the bias in future work. We estimate the needed quantities as follows. For convenience, let $v_i = i$ and $v_r = v - i$ for the numbers of passengers inspected and released respectively.

$$\begin{aligned}
\hat{\mu}_x &= b + v_i \times \frac{y_i}{n_i} \\
\hat{\sigma}_x^2 &= \frac{y_i}{n_i} \times \frac{n_i - y_i}{n_i} \times \frac{v_i^2}{n_i} \\
\hat{\mu}_y &= v_r \times \frac{y_r}{n_r} \\
\hat{\sigma}_y^2 &= \frac{y_r}{n_r} \times \frac{n_r - y_r}{n_r} \times \frac{v_r^2}{n_r}
\end{aligned} \tag{B.14}$$

The following example illustrates the calculation.

```
> v = 10000
> v_i = i = 3000
> v_r = v - v_i
> b = 30
> n_i = 100
> y_i = 5
> n_r = 300
> y_r = 5
```

We compute the quantities presented in equation B.14:

```
> mu.x = b + v_i * y_i / n_i
> s.x2 = y_i / n_i * (n_i - y_i) / n_i * v_i^2 / n_i
> mu.y = v_r * y_r / n_r
> s.y2 = y_r / n_r * (n_r - y_r) / n_r * v_r^2 / n_r
```

Then the estimated interval for $\hat{r}$ is

```
> (int.r = (mu.x * mu.y + c(1,-1) * 1.96 *
+             sqrt(mu.x^2 * s.y2 + mu.y^2 * s.x2 - 1.96^2 * s.x2 * s.y2)) /
+          (mu.x^2 - 1.96^2 * s.x2))

[1] 2.54697732 0.08177522
```

We can check it with the following arguably clearer code, which computes equations B.10 and B.11.

```
> p.a = mu.x^2 - 1.96^2 * s.x2
> p.b = -2 * mu.x * mu.y
> p.c = mu.y^2 - 1.96^2 * s.y2
> (int.r = (-p.b + c(1,-1) * sqrt(p.b^2 - 4 * p.a * p.c)) / (2 * p.a))

[1] 2.54697732 0.08177522
```

Then the point and interval estimates for NCE_s are as follows.

```
> (e.hat = (b + i * y_i / n_i) /
+           (b + i * y_i / n_i + (v - i) * y_r / n_r))

[1] 0.6067416

> (int.e = 1 / (1 + int.r))

[1] 0.2819302 0.9244065
```

B.2 Delta Method

We can also tackle the problem of obtaining an estimate of the confidence interval using an approximation, as follows. Recall that

$$\begin{aligned} NCE_s &= \frac{b + \hat{l}_i}{b + \hat{l}_i + \hat{l}_r} \\ &= \frac{1}{1 + \frac{\hat{l}_r}{b + \hat{l}_i}} \\ &= \frac{1}{1 + \hat{r}} \end{aligned} \tag{B.15}$$

where $\hat{l}_i$ is the estimated number of non-compliant units that were not detected by inspection, $\hat{l}_r$ is the estimated number of non-compliant units that were not detected by screening, b is the number of non-compliant units detected by inspection, and

$$\hat{r} = \frac{\hat{l}_r}{b + \hat{l}_i} = \frac{y}{x} \tag{B.16}$$

is now the ratio of the estimates of two independent random variables.

We can now reduce the problem to one of finding a confidence interval for $\hat{r}$ and then back-transforming the interval using $NCE_s = 1/(1 + \hat{r})$. Finding the interval for $\hat{r}$ can be done by estimating the standard error for the ratio and assuming that the sampling distribution for the statistic is approximately normally distributed. We need to find the standard error for

$$\hat{r} = \frac{v_r \times \frac{y_r}{n_r}}{b + v_i \times \frac{y_i}{n_i}} \tag{B.17}$$

where the quantities are as defined above.

We can use the Delta method to determine a first-order approximation to the standard error of a ratio of two independent random variables. Briefly, the Delta method involves taking a Taylor Series Expansion of the function about the estimate, and ignoring the higher-order terms on the grounds that their contribution is negligible. The outcome is the following expression for the variance of the ratio of two random variables, $y = \hat{l}_r$ and $x = b + \hat{l}_i$:

$$\sigma_r^2 = \text{Var}\left(\frac{y}{x}\right) = \left(\frac{\mu_y^2}{\mu_x^4}\right)\sigma_x^2 + \left(\frac{1}{\mu_x^2}\right)\sigma_y^2 - 2\left(\frac{\mu_y}{\mu_x^3}\right)\rho\sigma_x\sigma_y \tag{B.18}$$

where ρ is the correlation between b and l_r, which we assume to be 0, and $\sigma_x = \sigma_{\hat{l}_i}$ because b is a known constant.

So, the estimate of the variance of the ratio of the estimates is as equation B.18, with the pieces identified in Equation B.14, as follows.

$$\hat{\sigma}_{\hat{r}} = \hat{r}\sqrt{\frac{\hat{\sigma}_y^2}{y^2} + \frac{\hat{\sigma}_x^2}{x^2}} \tag{B.19}$$

So we use equation B.19 as an estimate of the standard error of the ratio, $\hat{\sigma}_{\hat{r}}$, and compute

$$\begin{aligned} NCE_s^u &= \frac{1}{1 + (\hat{r} - 1.96 \times \hat{\sigma}_{\hat{r}})} \\ NCE_s^l &= \frac{1}{1 + (\hat{r} + 1.96 \times \hat{\sigma}_{\hat{r}})} \end{aligned} \tag{B.20}$$

This interval estimate is acceptable so long as $b + \hat{l}_i > \hat{\sigma}_{b+\hat{l}_i}$ (**?**).

We provide a numerical example using the same data as before. Recall that $x = b + \hat{l}_i$ and $y = \hat{l}_r$. First, we check the condition.

```
> mu.x       # This is b + l_i

[1] 180

> sqrt(s.x2) # This is s_{b + l_i}

[1] 65.38348
```

This doesn't look great for our purposes — the mean is less than three standard deviations away from zero.

```
> hat.r = ((v - i) * y_r / n_r) / (b + i * y_i / n_i)
> s.r = hat.r * sqrt(s.y2/mu.y^2 + s.x2/mu.x^2)
> int.r = hat.r + c(-1,1) * 1.96 * s.r

> (hat.e = 1 / (1 + hat.r))

[1] 0.6067416

> (int.e = 1 / (1 + int.r))[2:1]

[1] 0.4208076 1.0870590
```

Both estimates of the confidence interval are easy to compute in a spreadsheet. Unfortunately, as an estimate they have some poor properties when the sample size is small, for example, the limits can be negative, or greater than 1. Under such circumstances we would take the limit to be 0 or 1 respectively, but it may be preferable to compute the interval using Monte-Carlo simulation, which is beyond the scope of this report.