cebra
Centre of Excellence for
Biosecurity Risk Analysis

CEBRA 20NZ01: Design a Statistically Valid Pathway Slippage Audit System

Final Report

Andrew Robinson, CEBRA, The University of Melbourne

Nikola Merrilees, The Ministry for Primary Industries, New Zealand

Michael Ormsby, The Ministry for Primary Industries, New Zealand

April 8, 2022

Contents

Table of Contents **1**

List of Tables **2**

List of Figures **2**

1 Introduction **5**
1.1 Background . . . 5
1.2 Intended outcomes . . . 5
1.3 Process . . . 5

2 Case study: unaccompanied personal effects **6**
2.1 Background . . . 6
2.2 MPI UPE risk management system . . . 6
2.3 Australian UPE system (for comparison) . . . 7
2.4 UPE data from MPI . . . 8

3 Slippage survey design **12**
3.1 Purpose of slippage measurement . . . 12
3.2 Options for slippage survey intervention approaches . . . 12
3.3 Slippage survey sample design . . . 12
3.4 Proposed extended analysis (Technical) . . . 13
3.5 Approach to sample-size determination . . . 14
3.5.1 Background . . . 14
3.5.2 Method . . . 14
3.5.3 Caveats . . . 15
3.5.4 Structured expert judgement exercise . . . 15
3.5.5 Results and discussion . . . 15
3.5.6 Choosing a sample size . . . 18

4 Recommendations & reflections **20**
4.1 Recommended sample design: Option 1 . . . 20
4.2 Recommended sample design: Option 2 . . . 20
4.3 Other considerations . . . 21
4.4 Future slippage survey design . . . 21

Bibliography **25**

A Initiation Workshop Notes **26**

B SEJ Workshop Participant notes **31**
B.1 Round 1 notes . . . 31
B.1.1 Question 1: upper risk threshold . . . 31
B.1.2 Question 2: lower risk threshold . . . 31
B.1.3 Question 3: low-risk goods . . . 32
B.1.4 Question 4: high-risk goods . . . 32
B.2 Workshop Minutes . . . 33
B.2.1 Thresholds . . . 33

B.2.2 HR pathway . . . 33
B.2.3 LR pathway . . . 34
B.3 Follow-up notes . . . 35

C SEJ Briefing notes **36**
C.1 On expert judgement . . . 36
C.2 Structured expert elicitation . . . 37
C.3 IDEA: Modified Delphi-style elicitation . . . 38
C.4 Description of the IDEA protocol . . . 38
C.5 Applications of expert elicitation . . . 39

D Technical details **40**

E Simulation study **41**
E.1 Preliminary work . . . 41
E.1.1 Ratio estimator . . . 42
E.1.2 Slippage only estimator . . . 42
E.1.3 Conditional estimator . . . 42
E.2 Targeting efficiency estimate . . . 42
E.3 Examples of HR calculation . . . 42
E.4 Simulation design . . . 43
E.5 Outcome reporting . . . 43
E.6 Results . . . 44
E.6.1 Pathway contamination estimate . . . 44
E.6.2 Targeting efficiency estimate . . . 44
E.6.3 BAU inspection efficiency estimate . . . 47

List of Tables

1 Average annual arrival counts, inspection rates, and action rates per inspection for Unaccompanied Personal Effects, 2018–2020, by country. . . . 11
2 Four options for the approach to measuring slippage in conducting slippage surveys for UPE's. . . . 13
3 Worked example for candidate estimators for high-risk stratum contamination weight. 43

List of Figures

1 The annual count of UPE consignments arrivals, by country of origin. . . . 9
2 The annual proportion of UPE consignments that were inspected, by country of origin 10
3 The annual proportion of UPE consignments that were treated, destroyed, or re-shipped if they were inspected, by country of origin. . . . 10
4 Unaccompanied personal effects overall contamination prediction and risk thresholds by expert. . . . 16
5 Unaccompanied personal effects contamination by risk stratum for each expert. . . . 17
6 Final-round country-specific expert-elicited consignment contamination rates for UPE. 19
7 Example of elicited distribution. . . . 40
8 Simulated 95% confidence interval coverage of the contamination estimate for three estimators. . . . 44

9 Simulated 95% confidence interval width of the contamination estimate for three estimators, in the units of the contamination rate. 45
10 Simulated root mean squared error of the contamination estimate for three estimators. 45
11 Simulated 95% confidence interval coverage of the targeting efficiency estimate for the approximate estimator. 46
12 Simulated 95% confidence interval width of the targeting efficiency estimate for the approximate estimator. 46
13 Simulated root mean squared error of the targeting efficiency estimate for the approximate estimator. 47
14 Simulated 95% confidence interval coverage of the inspection efficiency estimate for the approximate estimator. 48
15 Simulated 95% confidence interval width of the inspection efficiency estimate for the approximate estimator. 48
16 Simulated root mean squared error of the inspection efficiency estimate for the approximate estimator. 49

Executive Summary

CEBRA was asked by the Ministry of Primary Industries (MPI) to design a statistically valid and efficient slippage audit for the unaccompanied personal effects pathway. The process involved workshops with MPI personnel, interviews of Department of Agriculture, Water and the Environment (DAWE) personnel, and data analysis.

The recommended sample design is as follows. The following approach should be applied for one year or until a sufficient sample has been collected, and then the results analysed and conclusions drawn as to whether to continue, stop or change the slippage survey, and whether to retain or change the policies around border intervention for unaccompanied personal effects (UPE).

- Randomly select approximately 1000 apparently low-risk consignments; inspect according to the DAWE protocol (detailed in Section 2.3). This is approximately an 8% sample at the 2020 arrival rate.
- Randomly select approximately 500 apparently high-risk consignments; inspect according to the MPI protocol and then the DAWE protocol, that is, examine the declared items first, record the outcome, and then apply the DAWE protocol and record the DAWE outcome separately. This is approximately an 8% sample at the 2020 arrival rate.

NB: consignments inspected for the slippage survey should be flagged as slippage consignments in the database to enable efficient data analysis.

Upon collection of the data it should be analysed to estimate (i) the actual contamination rate of the low-risk pathway that is currently released on documents, and (ii) the comparative effectiveness of the MPI process and the DAWE process for inspecting UPE.

At that point MPI should reconsider both the regular intervention and the slippage survey requirements for this pathway in the light of these results.

Future slippage survey exercises may be able to apply a streamlined version of the approach that is documented in this report.

Acknowledgments

The authors are very grateful to the following experts for their patience and thoughtful efforts: Colin Bond, Jenna Ellis, Heidi Lovatt, Karlis Skuja, Manmeet (Mani) Sidhu, Danielle (Dani) Smith, and Fiona Stewart; the contributors to the initial workshop: Mervyn Alexander, Allan Burne, Chris Denny, Jenna Guest, Deepak Kumar Gupta, Francie Olliver, Evelyn Pleydell, Karen Pugh, Stu Rawnsley, Andrew Spelman, and Sina Waghorn; and furthermore Andrew Carson, Udaya Senarath, Nazre Sobhan, and Stephen Larcombe from the Department of Agriculture, Water and the Environment for kindly providing the information about the department's processes and practices for managing the biosecurity risk presented by UPE. Rob Cannon and David Fox provided invaluable review perspectives that strongly encouraged a rethink of the structure and presentation; the authors are grateful for their forthright candour.

1 Introduction

1.1 Background

A number of risk assessments by the Biosecurity Intelligence Team of MPI would benefit from a better understanding of slippage of actionable biosecurity risk material (ARM) across the various pathways, specifically from knowing what biosecurity risk is not being managed by the pre-border and border mitigations. Border Clearance Service do run slippage audits, however, we are not aware of a formal statistical framework under which these audits are designed and run.

1.2 Intended outcomes

CEBRA will design a statistically valid and efficient slippage audit for at least one pathway.

1.3 Process

1. CEBRA ran a stakeholder workshop (see Appendix A) to
   (a) identify pathways on which material slippage is felt likely to occur
   (b) determine breadth of potential data sources
   (c) establish decisions that would arise from slippage information
   (d) determine loss arising from different errors for decisions, and
   (e) prioritize pathways for analysis, aiming to choose exemplar pathways
2. MPI provided data to CEBRA on the nominated pathway (Unaccompanied personal effects, UPE), including volumes and types of risk goods, infestation rates (actual and target), etc. The data are summarised in Section 2.4.
3. CEBRA ran a structured expert judgement (SEJ, also called expert elicitation) exercise to estimate
   (a) the rate at which the apparently low-risk stratum of the UPE pathway could be carrying ARM, along with decision thresholds, in other words levels of contamination at which the approach to management of the biosecurity risk of the pathway should change; and
   (b) the rate at which the apparently high-risk stratum of the UPE pathway could be carrying undetected ARM, along with decision thresholds;

   (Section 3.5.4, and Appendix B).
4. CEBRA applied principles of sample design for decision-making to determine sampling approach and sample size across the UPE pathway (Section 4).
5. CEBRA carried out a simulation study to assess the informativeness of various versions of the sample design, and to verify the recommended approach and estimators.

2 Case study: unaccompanied personal effects

2.1 Background

UPE, or unaccompanied personal effects, are a class of non-commercial cargo. The UPE pathway is difficult to manage for biosecurity risk because documentation accompanying UPE consignments lacks economic tariff information, which is used as a basis for prescribing biosecurity risk management in commercial cargo pathways. Therefore there is very little information that can be used for profiling to provide insight into the pathway risk structure.

When carried out, inspection of UPE in Australia and New Zealand is manual, meaning that boxes are opened and the contents verified by a biosecurity officer or an approved person. The lack of profiling information creates more scope for initiative on the part of the inspector, who may learn more from opening the first box than from the accompanying paperwork. In this sense the management of biosecurity risk in UPE is more similar to that for international passengers than to that for cargo.

The primary drivers of biosecurity risk in the UPE pathway are (i) the items that are not inspected: misdeclared or undeclared items, and also items that unexpectedly carry or accompany risk (target slippage); and (ii) items that are inspected but for which some risk remains after inspection (inspection slippage).

The purpose of the proposed slippage survey is to assess the magnitude of target slippage; assessing the efficacy of inspection is not in scope. Nonetheless, the following design process includes allowance for estimating the inspection efficiency and offers options for implementation.

2.2 MPI UPE risk management system

Cargo interventions are imposed by the Ministry for Primary Industries (MPI) based on profiling using the economic tariff code from the Trade Single Window interface (TSW); there is a dedicated personal effects tariff, namely 98.05 Household or other effects, excluding aircraft, ships, small ships (as defined in concession reference number 80 in Part 2), and motor vehicles (including motorcycles) that are not intended for any other person, or for gift, sale, or exchange

9805.00.15 00K — Imported by a person to whom all of subparagraphs (i) to (iii) of paragraph (e) of concession reference number 80 in Part 2 apply

9805.00.89 00C — Imported by any other person.

Presently there is no regulatory framework for the management of UPE biosecurity risk, which are similar in this way to international passengers, although development is underway for the latter.

As a summary of activity on the UPE pathway, 1267 lines of UPE were created in October 2020. The total number of lines of interest to MPI was 21000 (excluding containers); UPE comprises 5–8% of cargo (11M total, $\sim 7\%$ of MPI interest; about 15K UPE per year, although this is not reflected in Figure 1). Once inside the border, the UPE go anywhere in the country, but mostly to Auckland, Tauranga, and Christchurch.

The risk of the UPE consignment is assessed based on a score computed from the bill of lading, manifest, and shipping documentation, which is a joint form also including Customs information, filled in by the importer with where they have been and what they think is risky. The listed items are allocated a score using an MPI rubric, and the score is used to stratify the consignment into one of two strata, namely: *always inspect* and *never inspect*. The consignment documentation is often handwritten and vague, e.g., "Box".

A total score > 5 results in referring for inspection everything that would usually be inspected. Consignments with scores ≤ 5 are released on documentation. There are also certain articles that result in mandatory inspection. Such inspection involves referral of all applicable items, not just the automatic inspection articles. Consignments that score < 5 are rarely inspected. We will refer to this as Business As Usual (BAU) intervention.

The items sent for inspection are recorded, but the reason for inspection is not, so we can only infer whether the consignment score was > 5 or contained must-inspect items by a process of elimination.

Inspections are cost-recovered at $100 per hour, with a minimum rate of 15 minutes plus 15 for paperwork; there is also a cargo levy on every Customs entry.

Inspection outcomes are recorded in a single UPE line, the high-risk articles within the consignment that are identified for inspection are concatenated in the database as a single comment. Outcomes are hard to identify. Furthermore, the low-risk cargo information is recorded in scanned pdfs. In order to extract inspection information (e.g., items inspected) these pdfs would need to be opened and read, which would be considerable manual work.

There is substantial uncertainty about the underlying biosecurity risk in the pathway. Border inspection has a low interception rate (see Section 2.4), but the data capture has unknown fidelity. The inspectors are pragmatic, so for example they might find grass clippings on a lawn mower and sweep them off to manage the risk, rather than formally recording the interception. Consequently, according to the interception rate, the pathway seems relatively clean but this is uncertain.

If there is an interception, then the response depends on the type of interception; for example MPI might clean or treat the article using an approved treatment provider. MPI may identify captured specimens down to species level, and may call importer and request permission to destroy. MPI may send a Biosecurity Authority/Clearance Certificate to document what needs to happen to the cargo. The outcome depends on whether the contaminant can be destroyed and on the officer and importer.

A quality assurance (QA) desktop exercise checks the border decision-making of 1% of all cargo applications (not just UPE). The QA re-screen goes through the consignment's paperwork again, checks data entry and charges, and so on. To date, the QA exercise shows a 96% agreement with border actions.

The implications of the outcome of a slippage survey are as follows. MPI might need to create or reinstate a UPE standard if the pathway were discovered to be of unexpectedly high risk. Also there could be detection of some high-risk pests: for example, brown marmorated stink bug (BMSB), Asian gypsy moth (AGM), or fruit fly. If necessary, MPI policies might build on those for accompanied luggage. Risk factors for UPE are held to be more strongly indicated by the types of cargo e.g., outdoor furniture, BBQ, lawnmower, than by the country from which the importer has traveled, or their personal details. However, there are some elements in both types of profiling that are similar such as care taken in completing documentation, understanding of English, etc.

2.3 Australian UPE system (for comparison)

This section provides a brief description of the Australian Federal Department of Agriculture, Water and the Environment (DAWE) system for managing the ingress and biosecurity risk presented by the UPE pathway[1].

All imported cargo is reported to the Australian government through Australian Border Force's Integrated Cargo System (ICS). Unaccompanied Personal Effects (UPEs) are reported on a cargo report (made by the carrier or freight forwarder) and typically declared as *personal effects* on this

[1]This information was kindly provided by Andrew Carson, Nazre Sobhan, and Udaya Senarath of DAWE.

report at the same time. If so declared, then the consignment is electronically held for department assessment.

The Biosecurity Import Conditions database (BICON), which is the department's online database of biosecurity risk management requirements, specifies the department's conditions under which personal effects can be imported. The importer must complete:

- an Unaccompanied Personal Effects Statement (B534 form), which contains details like origin, destination, goods, and high-risk factors such as timber, animal products, etc.
- a packing list, which is usually hand-written and can be quite generic e.g., Bedroom 1, Bedroom 2, kitchen, table, chairs, lawn mower etc.

Department staff assess these documents (B534 and packing list) and check BICON and then either release the consignment or direct it for inspection. If the consignment is large (> 100 kg OR $> 1\text{m}^3$) then it is always directed for inspection. If the consignment is smaller and the documents justify an assessment of low risk then the consignment is released. About 10% of all consignments are released without inspection.

Inspection is carried out at the approved arrangement (AA) site, with most UPE inspections being conducted at professional removalist company AA sites. The packing list is used to guide the inspection, for example focusing on the kitchen (looking for spices), the garden shed (tools), timber furniture and so on. AA site staff will unpack the container and identify likely higher-risk articles/boxes from the packing list and set them aside. The officer will check the packing list and identify and request any further articles of interest.

The initial inspection can comprise any proportion of the consignment based on the documents provided e.g., 10% to 100%. After the potential risk boxes/items are inspected, staff inspect a further random 10% of the consignment. If the officer deems it necessary, then further articles may be inspected. We will refer to this inspection protocol as the DAWE protocol.

There are several databases in which UPE documentation may be recorded, including self-assessed clearances (SAC). There are approximately 50,000 arrivals of UPE accompanied by a B534 per year. Biosecurity risk interceptions are recorded in the Agriculture Import Management System (AIMS) and the Incidents Client database.

Further information is available at the department's webpage on moving to Australia or importing personal effects/household goods[2] and the Australian Border Force website on unaccompanied personal effects.[3]

2.4 UPE data from MPI

This section provides a high-level summary of the UPE inspection data held by MPI. The purpose of the analysis was to determine the volume of consignments in the pathway and to describe the rates at which inspections are undertaken and ARM is detected. The dataset comprised 159315 entries, with dates ranging from Jan 1 2009 until Jan 31 2021. There were 20561 entries in 2020, 7108 of which were inspected (roughly a third).

Figure 1 shows the annual count of UPEs by country, focusing on the last decade (2009–2020) and including only the top eight countries in terms of total arrivals. The data show a peculiar and (relatively) substantial dip in the arrival rates from 2011 to 2019 that is not mirrored in immigration records (not shown here) and therefore has no easy explanation. The dip is particularly marked in arrivals from Australia and UK, but is not much smaller relatively speaking in the data from

[2] `https://www.agriculture.gov.au/travelling/moving-immigrating`

[3] `https://www.abf.gov.au/entering-and-leaving-australia/moving-to-australia/upe`

the other countries. MPI considers this dip is likely to be from an undocumented infelicity in data capture or storage and is very unlikely to reflect a genuine pattern in the arrival of UPE.

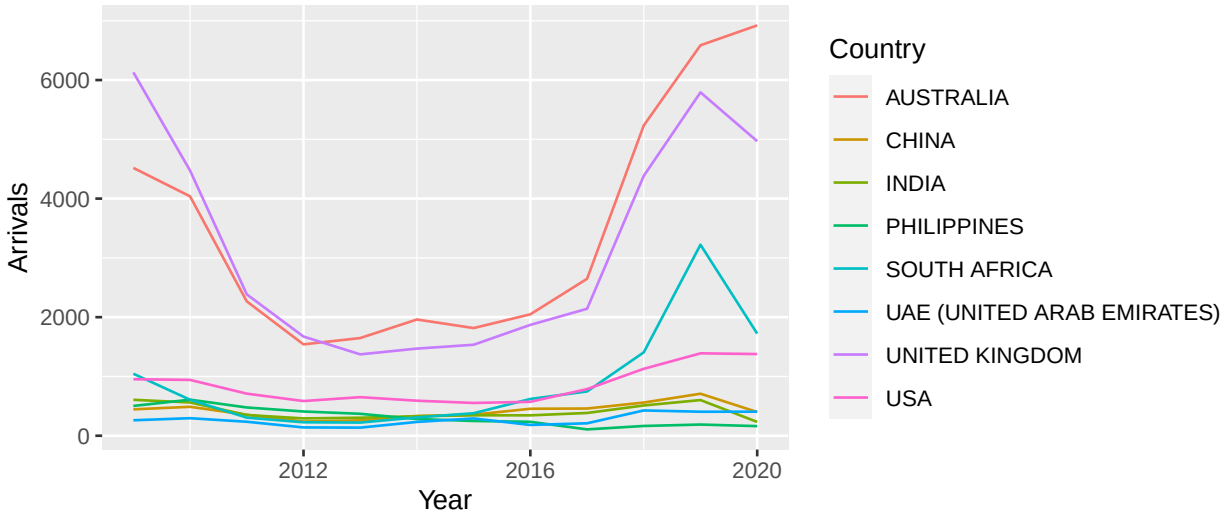

Figure 1: The annual count of UPE consignments arrivals, by country of origin.

Figure 2 provides a summary of the proportion of consignments that were inspected, or more accurately, the proportion of consignments for which inspection outcomes were recorded. These inspection rates show a similar drop as is seen in Figure 1, but cannot be explained simply by a reduction in consignment counts as the statistic being plotted is proportion, rather than count, of consignments that are inspected. It is possible that more consignments were inspected than are recorded, and these records were stored in a separate database. More recently, the data show a high proportion of inspections for consignments that arrived from South Africa, USA, Australia and UAE, and about half the proportional rate for consignments from UK, and smaller proportions still for India, the Philippines, and China.

The more complicated story of action proportion and inspection proportion combined, per country per year, is provided in Figure 3. We see that there is not much evidence of a trend in the action proportion for any country. The strange dips noted in Figures 1 and 2 are obvious in the relative sizes of the points, which reflect the number of inspections: the large points are at the start and the end of the series, and there are none in the middle.

Overall the USA has the highest historical action proportion among the countries for which data have been plotted, followed by Australia, but this may be explained in a number of ways, for instance the consignments from these countries may be inherently riskier, or the documentation may be more accurate. Note that the inspection proportion is higher for consignments from those two countries (Figure 2). The overall action proportion for consignments from the UK and South Africa follow, and the action rate is lowest for consignments from China and the Philippines.

The overall inspection proportions and action proportions in the 2019 data by country, for the top 31 countries (and Other), are provided in Table 1. The predominant source for UPE consignments was Australia and the UK. Inspection proportions were high for Tonga, South Africa, USA, Germany, and Australia. Action proportions were high for Tonga, France, Italy and Samoa. It is important to keep in mind that these statistics do not, in and of themselves, provide any indication as to whether more intervention on the apparently high-risk pathways would result in a reduction in biosecurity risk. In order to draw such a conclusion, a slippage survey must be

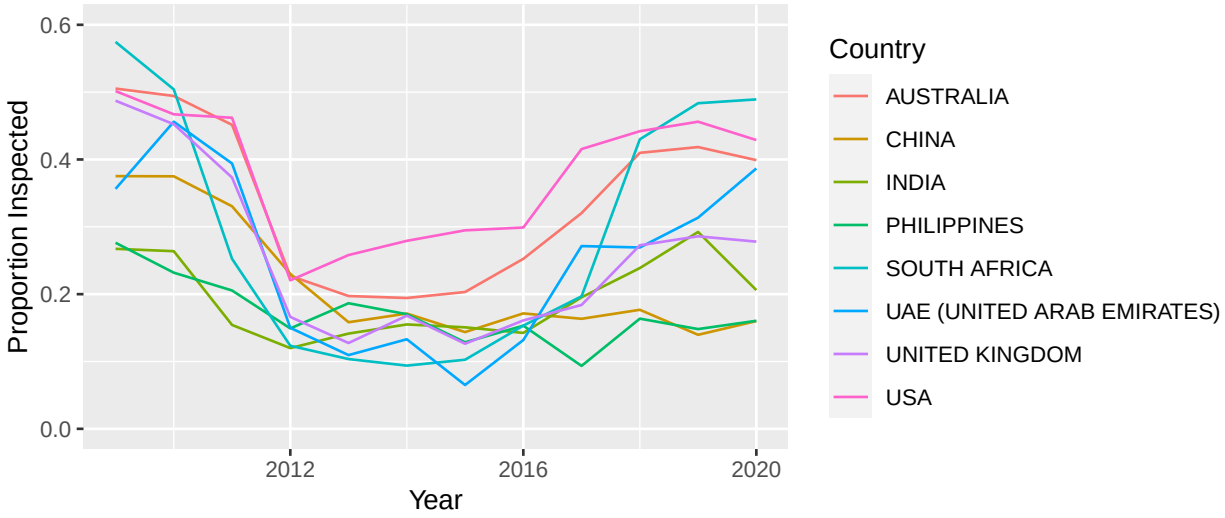

Figure 2: The annual proportion of UPE consignments that were inspected, by country of origin. Inspection policies differed over the timespan covered by the data.

performed in order to assess the contamination rate in the apparently low-risk stratum.

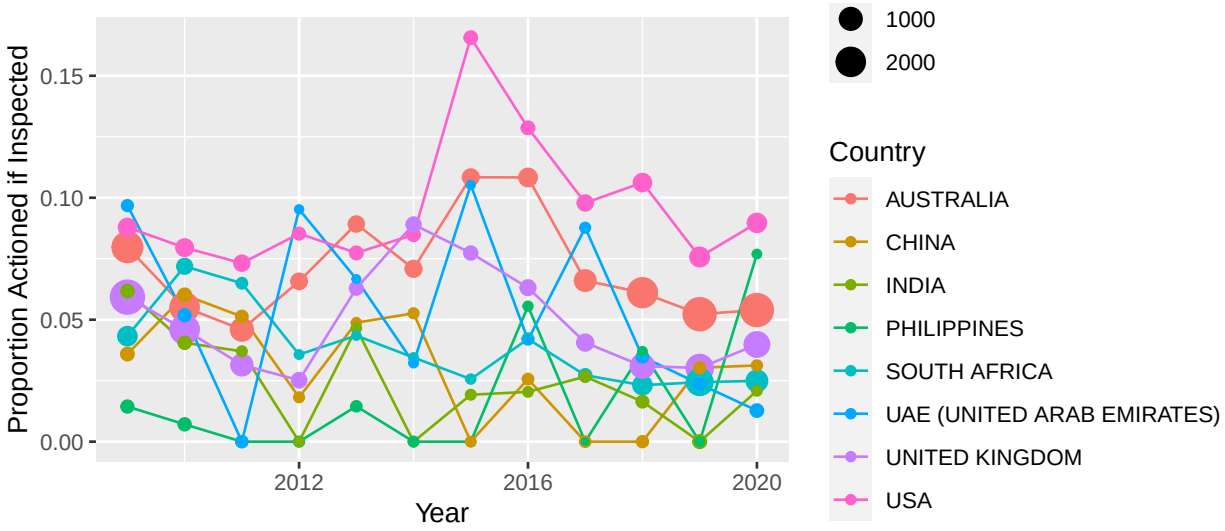

Figure 3: The annual proportion of UPE consignments that were treated, destroyed, or reshipped if they were inspected, by country of origin. Point size reflects the number inspected.

Table 1: Average arrival counts, inspection rates, and action rates per inspection for Unaccompanied Personal Effects, 2018–2020, by country. *Action* reports the proportion of the inspected consignments that were treated, destroyed, or reshipped. Countries with fewer than 50 entries in 2019 are collected in *OTHER*.

Country	Total	Inspected	Action
AUSTRALIA	6247	0.41	0.064
UNITED KINGDOM	5048	0.28	0.036
SOUTH AFRICA	2117	0.47	0.025
USA	1299	0.44	0.103
OTHER	909	0.33	0.113
CHINA	556	0.16	0.034
INDIA	449	0.26	0.020
UAE (UNITED ARAB EMIRATES)	413	0.32	0.025
SINGAPORE	410	0.28	0.026
HONG KONG	394	0.25	0.007
KOREA - SOUTH	367	0.27	0.023
CANADA	363	0.34	0.053
NETHERLANDS	216	0.32	0.073
FIJI	188	0.32	0.094
PHILIPPINES	172	0.16	0.099
GERMANY FEDERAL REPUBLIC OF	168	0.43	0.078
JAPAN	152	0.25	0.053
FRANCE	137	0.33	0.148
TONGA	130	0.59	0.491
MALAYSIA	118	0.29	0.067
THAILAND	114	0.24	0.049
IRAN ISLAMIC REPUBLIC OF	104	0.26	0.012
IRELAND	103	0.19	0.083
SAUDI ARABIA	90	0.39	0.010
ITALY	82	0.17	0.186
SAMOA	79	0.38	0.189
INDONESIA	73	0.28	0.065
QATAR	62	0.25	0.000
SRI LANKA	55	0.35	0.017
SPAIN	53	0.20	0.031
PAKISTAN	46	0.06	0.000
BELGIUM	45	0.45	0.067

3 Slippage survey design

This section details the process of developing a slippage survey. Briefly, the process is to (i) determine the purpose of measuring the slippage (Section 3.1), (ii) establish a mechanism for measuring the slippage (Section 3.2), (iii) select a suitable sample design (Section 3), and then (iv) determine a sample size (Section 3.5).

3.1 Purpose of slippage measurement

A slippage survey has two roles: first, to report on the slippage in the pathway with sufficient certainty as to be reasonably informative, and second: to provide motivation and evidence for any needed changes in the policy or practice of biosecurity risk management in the pathway. The former role suggests that the sample design should deliver a reasonably precise estimate across a wide range of circumstances. The latter usage implies that the pathway manager or risk owner must be able to identify thresholds at which remedial action should be taken: this could mean, for example, that profiles be amended, or intervention policies updated, and so on.

3.2 Options for slippage survey intervention approaches

The options for slippage survey intervention can be established by brain-storming and/or interviews with subject-matter experts. Table 2 provides four options for the approach or mechanism to use for measuring UPE slippage, in order of decreasing priority, along with brief notes on their advantages and disadvantages.

For the purposes of the current exercise, we recommend *manual sampling*; the quality of the information gained in the context of the existing regulatory framework outweighs the benefits of the other approaches. Specifically, the slippage survey would be carried out in a way that is commensurate with the current targeted intervention regime, so risk items that are detected by the slippage survey are in scope for the current approach.

3.3 Slippage survey sample design

The sample design prescribes how the sample will be selected from the consignments on the pathway. The sample design therefore needs to be chosen by taking account of the characteristics of the pathway of interest. A suitable sample design will provide robust and efficient statistical estimates of process parameters of interest, such as the contamination rate, by nature of the statistical characteristics of the design itself (Cochran, 1977).

As noted above, MPI presently splits the stream of UPE into two strata using a risk score and the presence of must-inspect items. These two strata are treated differently because of their different levels of presumed risk, so it makes sense statistically as well as operationally to stratify the slippage survey in the same way.

Within the strata, other variables could potentially be used to further refine the slippage survey selection, such as consignment volume or weight, but the benefits of using these extra variables are speculative and their invocation would complicate implementation, so we do not consider them further here. Instead, we propose a data collection regime that would provide information about whether further efficiencies might be worth seeking.

Consequently we recommend a stratified design, following the current intervention approach, and the deployment of a simple random sample within each stratum. In terms of relative size, the strata are approximately 1/3 high-risk and 2/3 low-risk.

Table 2: Four options for the approach to measuring slippage in conducting slippage surveys for UPE's. Each approach is accompanied by pros, cons, and notes.

Approach	Pros	Cons	Notes
Manual sampling: select, open and empty random boxes of goods selected from random UPE cargo consignments.	The slippage inspection activity mirrors the authorised regulatory activity, so any slippage is theoretically detectable by inspection and therefore in scope from the point of view of estimating target slippage.	Expensive, time consuming	Following the Australian protocol, the inspection could use the packing list to guide inspection to focus on apparently risky parts of the consignment (e.g., kitchen, garden shed) and select random items thereafter up to 10%.
Detector Dog Unit: run the DDU over random UPE consignments as a screening measure.	Efficient. DDU reasonably portable.	Probably too many conflicting odours. Also, because the inspection method differs from the usual approach, slippage estimates will combine target and inspection slippage.	The DDU is already used in passenger and mail pathways. DDU could be trialed as an alternative risk management measure.
3-D tomography. Put smaller units through the Rapiscan machine at the airport or mail facility.	Efficient screening	There are procedural challenges with bringing goods back into the airport environment. The airport is not a transitional facility. Also, because the inspection method differs from the usual approach, slippage estimates will combine target and inspection slippage in equal quantities.	Auckland airport has a Rapiscan 3-D tomography scanning device that has algorithms tuned to detecting biological material for the purpose of scanning passenger baggage and mail articles.
CEF (Container examination facility) full container x-ray.	Efficient, and could be happening already	Detecting biosecurity risk within UPE could be difficult using whole-container X-ray, which is not designed for that purpose. Also, because the inspection method differs from the usual approach, slippage estimates will combine target and inspection slippage.	NZ Customs has CEFs — they do ask permission to move containers that are under MPI jurisdiction but no necessary information is provided. Customs will flag if x-raying anything that is of MPI interest.

3.4 Proposed extended analysis (Technical)

Following the simulation study documented in Appendix E, we recommend the use of the conditional estimator. The conditional estimator uses the information in the BAU inspection as follows: stratify the high-risk stratum into two sub-strata based on the outcome of the BAU inspection, detections and non-detections. Use the slippage survey in the non-detection sub-stratum to estimate the probability of a false negative, and then estimate the number of false negatives in the non-detection sub-stratum.

$$\tilde{p}_h = \frac{1}{N_h}\left(z_h + (N_h - z_h)\hat{\rho}\right) \tag{1}$$

where $\hat{\rho}$ is formed as follows: considering only the BAU negatives that were in the slippage sample, compute the proportion that were positive in slippage inspection.

That is, the estimated number of contaminated consignments is the sum of the contaminated consignments detected by BAU and the estimated number of contaminated consignments missed

by BAU, which is estimated as the product of the number of consignments that were not detected as contaminated and the probability of a false negative in BAU.

3.5 Approach to sample-size determination

3.5.1 Background

In order to set a sample size for a survey, it is essential to first determine the use or uses to which the information will be put. There is no point in undertaking a survey without a clear plan to use the outcome.

The usual approach to determining a sample size in advance of undertaking a survey is to specify a desired length of a confidence interval and to back-calculate the needed sample size that is required to achieve the confidence interval based on assumptions about the process. For example, ISPM 31 uses the Binomial or the Hypergeometric distributions to justify the well-known sampling rule of requiring about 600 sampling units to detect a contamination rate of 0.5% with 95% sensitivity (FAO, 2008).

A challenge with designing a survey for a Binomial random variable is that the sensitivity of the survey depends on the value that is being estimated.[4] Consequently, in order to predict the sensitivity of a future binomial estimate from a sample, paradoxically, we need to know the future estimate itself as well as the intended sample size.

In slippage survey situations, it may well be reasonable to design the survey assuming that the target or benchmark rate is 0, but that will not necessarily provide estimates that enable action with sufficient certainty, that is, that will identify whether the pathway risk management is acceptably within specifications. In order to be confident that the slippage survey is sufficiently valuable, we need to know (i) the risk thresholds, (ii) the loss from making different sorts of errors(!), and (iii) the true underlying contamination rate(!) Most of these are unknowable, so we must make do with the best information available.

3.5.2 Method

Our ultimate goal is to try to determine the true underlying contamination rate, the risk thresholds, and the loss arising from mis-classifying the pathway. To simplify the challenge we estimated the unknown underlying contamination rate using structured expert judgement (Section 3.5.4). At the same time, we elicited upper and lower thresholds for risk management, forming an interval estimate. To do this we asked the experts to identify a value above which they might conclude the pathway is too risky to be managed with the current intervention. We also asked the experts to each identify a value below which they might conclude the pathway is so safe as to require less intervention than is currently imposed. This exercise provides us with an estimate of the contamination rate, along with a measure of uncertainty, and an upper and a lower threshold for each expert. For the design we only considered the upper threshold.

We then computed a sample size that would be needed for each expert in order to have sufficient confidence of the location of the contamination rate relative to the thresholds *and* to be confident that the rate would be well estimated, using the standard approach. These approaches are detailed in Appendix D. In order to capture the expert's uncertainty we simulated from the implied probability distribution of the expert opinion of the contamination rate and reported the median and the 0.95 quantile. These distributions are then used to determine a reasonable sample size.

[4]The variance of the estimate of the mean of a binomial random variable, say p, arising from a sample of size n, is $p(1-p)/n$. This is in contrast to, for example, the variance of the estimate of the mean of a normal random variable, which is independent of the mean.

3.5.3 Caveats

As pointed out by a reviewer, a disadvantage with the approach of identifying thresholds and establishing a sample size that can be used to make a suitably powerful decision is that the approach focuses on how close the contamination rate is to a threshold and ignores the level of the contamination rate as an indicator of biosecurity risk. This throws up the possibility, for example, of devoting a very large number of samples to testing whether a pathway contamination rate is below 0.5%, when the true contamination rate might be so small as to provide operationally negligible information.

3.5.4 Structured expert judgement exercise

The objective of the SEJ exercise was to get the best possible idea of the biosecurity risk presented by the UPE pathway into New Zealand. Briefing materials were provided to the participants (see Appendix C), along with recent inspection history and four SEJ questions, namely

1. Imagine 1000 random UPE consignments coming from a range of countries to NZ. What is the lowest number of consignments that require intervention beyond standard inspection that would convince you that a different set of measures or policies are required to properly manage the biosecurity risk associated with UPE? That is, what is the lowest number that would say to you: this pathway is unacceptably risky and we have to do something more!

2. Imagine 1000 random UPE consignments coming from a range of countries to NZ. What is the highest number of consignments that require intervention beyond standard inspection that would comfort you that a lesser set of measures or policies are sufficient to manage the biosecurity risk associated with UPE? That is, what is the highest number that would say to you: this pathway is surprisingly safe! Let's do less here and focus our attention elsewhere!

3. Imagine 1000 random consignments of UPE from all countries. Their documentation corresponds to low-risk (that is, risk score ≤ 5 and no must-inspect items). How many of those 1000 random consignments will have actionable risk material?

4. Imagine 1000 random consignments of UPE from all countries. Their documentation corresponds to high-risk (that is, risk score > 5 or at least one must-inspect item). How many of those 1000 random consignments will have actionable risk material?

The experts provided initial interval estimates of the quantities (not shown here) and anonymised commentary on the questions as captured in Appendix B.1. These initial estimates were captured in graphical summaries similar to those in Section 3.5.5 and used as a basis for further discussion in the Phase 2 workshop, held online August 4 2021. The discussion from the workshop, and subsequent communications, are provided in Appendices B.2 and B.3 respectively.

3.5.5 Results and discussion

Figure 4 provides the overall results for the SEJ exercise. It is formatted as follows. Each expert has a dedicated row. Their opinion about the contamination rate of the pathway (across both low-risk and high-risk consignments) is represented as a violin on the plot (a violin plot is a symmetrical density plot). The point clouds represented by the violins are simulated from the base data stratified by declaration and country of origin (Figure 6) using the average volumes for 2017–2019 to provide stratum weights. Each expert also provided two thresholds, namely one below which the

pathway risk is lower than would justify the current measures, and one above which the pathway risk demands more stringent measures. These thresholds are captured as green and red dots on the graph respectively.

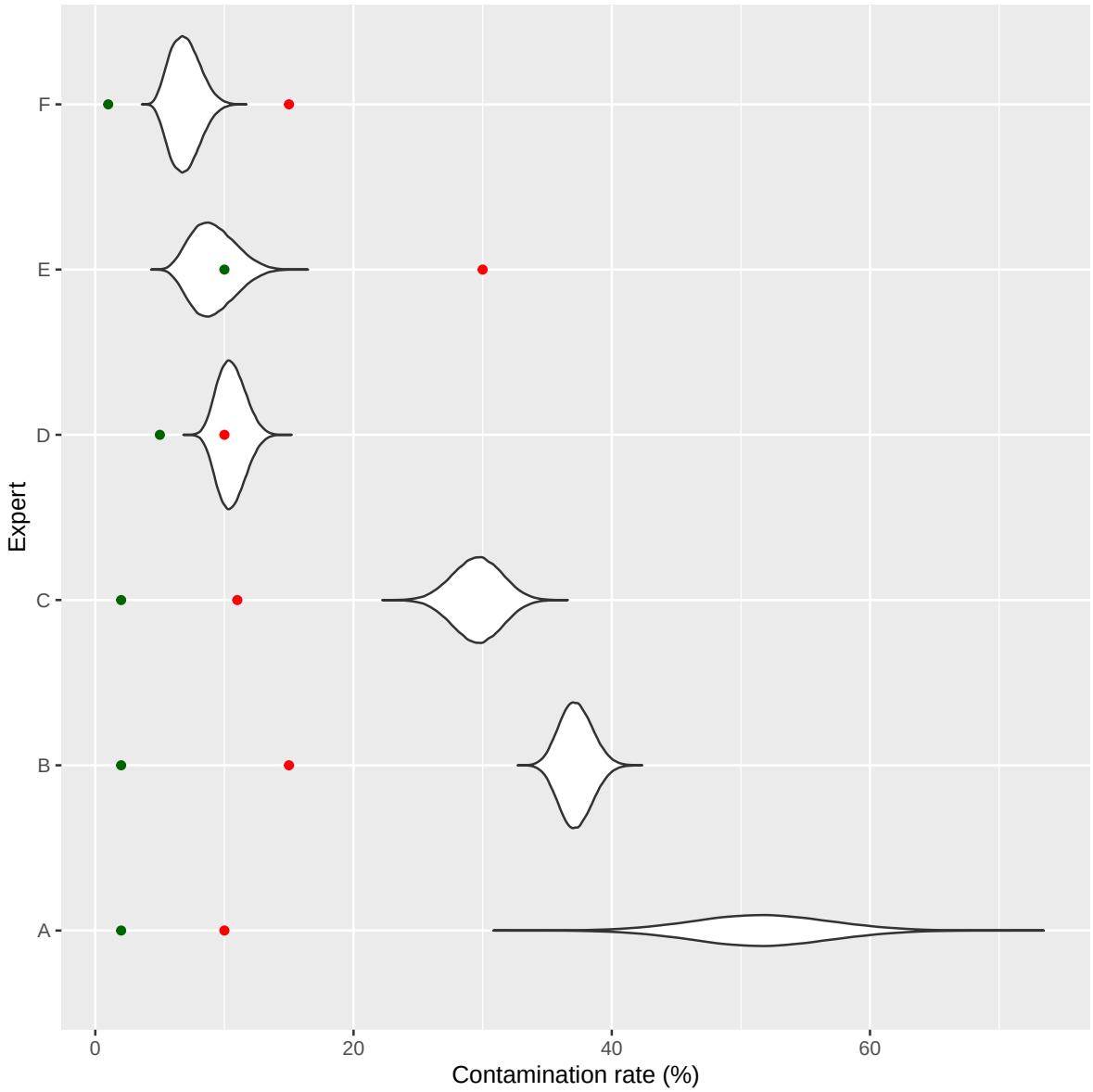

Figure 4: Unaccompanied personal effects overall contamination prediction and risk thresholds by expert. Each expert comprises a row, sorted by average contamination expectation. The violin represents the expert's uncertainty. The green dot is the lower-risk threshold, below which the pathway risk is lower than would justify the current measures, and the red dot is the higher-risk threshold, above which the pathway risk demands more stringent measures.

Figure 4 shows that the experts were split into two groups, namely relatively pessimistic (A, B, and C), who consider the pathway risk to be higher than their acceptable threshold, and relatively optimistic (D, E, and F), who consider the risk to be lower. Only one expert (E) considered that the risk might be low enough to contemplate less stringent measures; for four experts the risk was clearly higher than or quite close to their upper threshold. Expert F considered the risk to be within

their acceptable thresholds.

As noted above, the pathway risk is managed by the stratification into two categories of risk: low-risk consignments are those with a risk score less than 6 and no must-inspect items, for which documentation is assessed and then the consignment released, and high-risk consignments are those with a risk score greater than 5 or at least one must-inspect item, for which identified risk items are inspected manually. We asked the experts to distinguish between the contamination rates of the low-risk stratum (for which no inspections have been performed) and the high-risk stratum (which have been inspected). These results are captured in Figure 5.

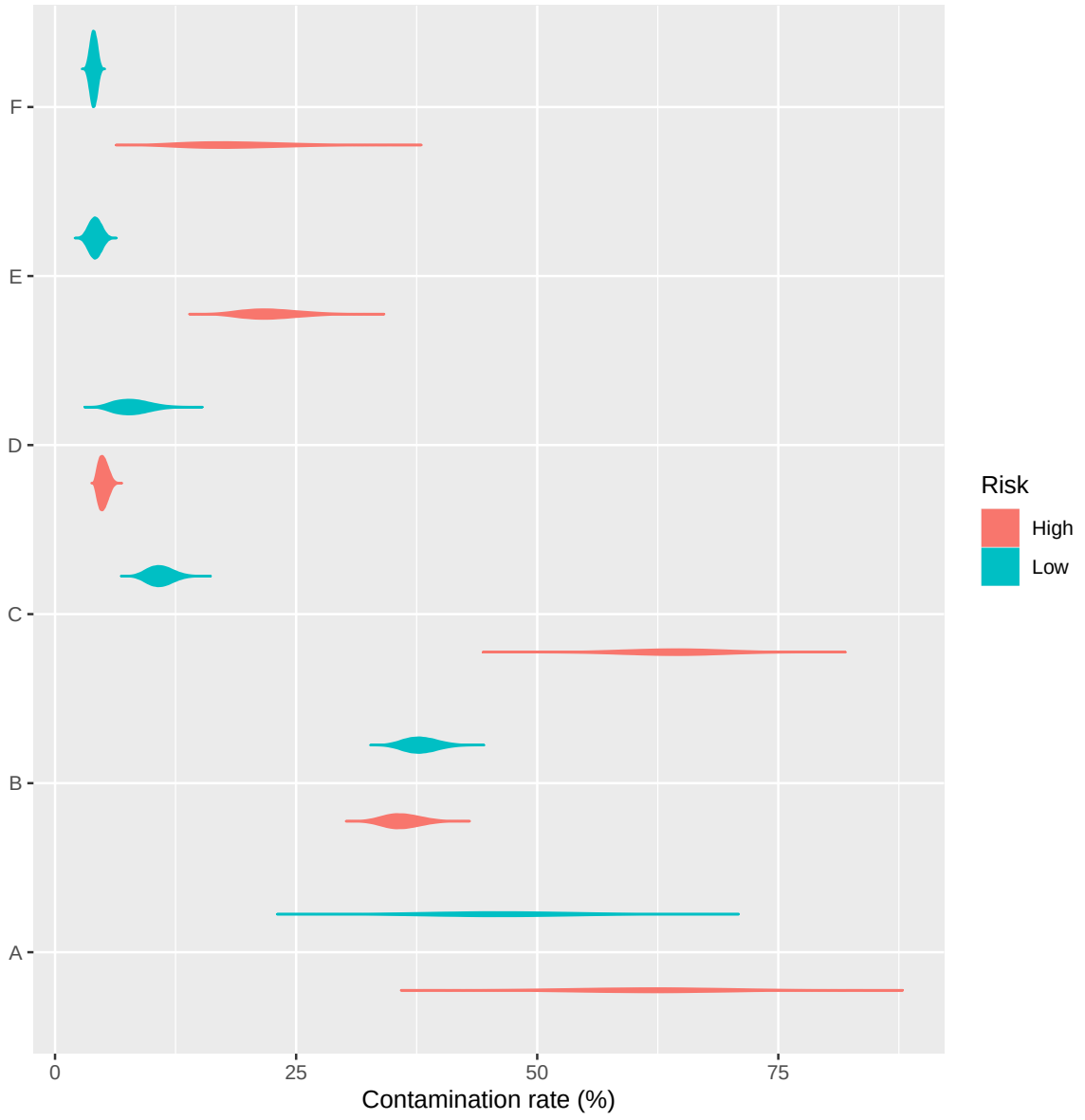

Figure 5: Unaccompanied personal effects contamination by risk stratum for each expert. Each expert comprises a row, sorted by average contamination expectation. The cyan density represents the lower-risk stratum, and the red density represents the higher-risk stratum.

Figure 5 compares the expert opinion on the contamination rate of the low and high strata by presenting both, paired up by expert. Only expert D accepted the interception history at face value;

the other five experts considered that the high-risk stratum has a substantially higher contamination rate than is detected during inspection (recall that the average interception rate is less than 10%, although substantially higher for some countries, e.g. Tonga, see Table 1). There was no substantial difference between the contamination rates of consignments within each stratum for five out of the six experts. Expert C was an exception, considering the high-risk stratum to have a substantially higher contamination rate than the low-risk stratum. Four experts (C–F) agreed that the low-risk stratum has a low contamination rate. Two experts (B and D) thought that the low-risk stratum will have a higher contamination rate than the higher-risk stratum, suggesting that stratification is deleterious for risk management.

Table 1 provides some evidence for the different rates at which ARM is detected during inspection for high-risk consignments arriving from different countries. Consequently, we offered the experts an opportunity to decide whether low- or high-risk consignments from any particular countries were more or less risky than the overall low- and high-risk consignments. The final round expert-elicited contamination results are presented in Figure 6.

Figure 6 shows that with few exceptions, the experts did not consider country to country differences to be of great importance. The discussion raised some possible explanations for systematic differences, for example that consignments from Australia, South Africa, the UK and USA were more likely to be complete households (and therefore to contain ARM), and countries from which importers are more likely to camp, hike, and garden, which are higher risk activities. However, the elicited differences between the countries was small relative to the uncertainty within the estimates (Figure 6).

The rule of analysis for SEJ exercises is to "aggregate late", so instead of taking the average of the high and low thresholds across the experts, and summing across the distributions and then normalising, we determine the sample size corresponding to each expert's judgement, and aggregate this distribution of needed sample sizes to arrive at the final design. We do this because if the experts have a different opinion on the relative risks of different types of ARM then there may be a correlation between how risky they consider the pathway to actually be and their thresholds for higher or lower risk management. A brief glance at Figure 4 shows no obvious relationship of the kind but we retain our plan regardless.

3.5.6 Choosing a sample size

The SEJ estimate of pathway contamination was used to calibrate a simulation study that is documented in Appendix E. On the basis of the simulation study, we recommend designing the slippage survey to achieve a confidence interval estimate width of about 0.01 on the contamination rate, so that if the rate were estimated at 0.07 (7%) then the confidence interval would be 0.065–0.075. This interval width can most conveniently and efficiently be achieved by taking 1000 sampling units from the low-risk stratum (say, 20 per week for a year) and 500 sampling units from the high-risk stratum (say, 10 per week for a year).

Figure 6: Final-round country-specific expert-elicited consignment contamination rates for UPE. Any country not listed is included in Overall. Best guesses are represented by dots, and upper and lower levels by vertical lines connected by a horizontal line. Expert responses are coloured consistently, to facilitate comparison. For example, Expert G in Overall corresponds to Expert A in India, and Expert C in South Africa, UK, and USA.

4 Recommendations & reflections

As a rule, a slippage survey should be undertaken randomly, and include enough sampled units that evidence of important contamination can be detected with reasonable probability and that the slippage rate can be estimated with reasonable precision, so that justifiable decisions can be made about risk management on the pathway. Furthermore, if there are important sub-populations in the pathway that are more or less likely to be risky, then a well-designed slippage survey of sufficient size should be designed to provide evidence for the utility and efficiency of bespoke intervention, perhaps by a risk-based algorithm such as one of the Continuous Sampling Plan family (Dodge, 1943; Robinson et al., 2012).

4.1 Recommended sample design: Option 1

The following approach should be applied for one year, and then the results analysed and conclusions drawn to continue, stop or change the slippage survey, and to retain or change the policies around border intervention for UPE.

- Randomly select approximately 1000 apparently low-risk consignments; inspect according to the MPI protocol (detailed in Section 2.2).

NB: consignments inspected for the slippage survey should be flagged as slippage consignments in the database to enable efficient data analysis.

Upon collection of the data it should be analysed to estimate the actual contamination rate of the low-risk pathway that is currently released on documents. The analysis should follow the specifications as laid out, for example, in ISPM 31 FAO (2008) or the accompanying spreadsheet.

At this point MPI should reconsider both the regular intervention and the slippage survey requirements in the light of these results.

4.2 Recommended sample design: Option 2

The second recommended sample design is as follows. It differs from the first in that it also estimates the effectiveness of the MPI inspection approach by comparing it with the more stringent DAWE approach. This approach is formally beyond the original remit of the project, but may nonetheless be preferred.

The following approach should be applied for one year, and then the results analysed and conclusions drawn to continue, stop or change the slippage survey, and to retain or change the policies around border intervention for UPE.

- Randomly select approximately 1000 apparently low-risk consignments; inspect according to the DAWE protocol (detailed in Section 2.3).
- Randomly select approximately 500 apparently high-risk consignments; inspect according to the MPI protocol and then the DAWE protocol, that is, examine the declared items first, record the outcome, and then apply the DAWE protocol and record the DAWE outcome separately.

NB: consignments inspected for the slippage survey should be flagged as slippage consignments in the database to enable efficient data analysis.

Upon collection of the data it should be analysed to estimate (i) the actual contamination rate of the low-risk pathway that is currently released on documents, and (ii) the comparative effectiveness

of the MPI process and the DAWE process for inspecting UPE. The analysis approach is provided in Section 3.4 and in the accompanying spreadsheet. Section E.3 also includes an approachable worked example.

At this point MPI should reconsider both the regular intervention and the slippage survey requirements in the light of these results.

Approximately a third of UPE consignments were inspected in 2020. The second proposed design allocates effort in proportion to the approximate volumes of activity in the low-risk and high-risk strata, and supports estimation of the target slippage and inspection slippage separately, enabling an improved estimate of the ARM contamination rate.

4.3 Other considerations

The statistical validity of the slippage survey depends on the nature of the sample-selection mechanism. We advocate simple random sampling within strata, that is: that within each stratum, the consignments to be inspected are selected randomly and independently of one another. This prescription can be considered of equal importance (if not more so) to selecting a sufficiently large sample. If the sample size is too low, but the sample is statistically valid, then the confidence interval estimate will be high but informative. If the sample size is reasonable but the sampling approach is not statistically valid then interpretation is difficult.

During the carriage of the project, a number of unanswered questions were raised that may be of interest to the analysis arising from the slippage survey.

1. Is manual inspection the most efficient way to manage the biosecurity risk presented by the UPE pathway? (see Section 3.2).

2. Other considerations may also be germane to the risk that is presented by the pathway.

   (a) Was the consignment packed by the importer or was it packed by a forwarder? This may affect the quality of the documentation.

   (b) The destination of the consignment within the country may affect the realised biosecurity risk

3. Is there any reason to believe that UPE would be different coming to NZ vs coming to Australia?

Finally, we reflect on the process of developing the slippage survey that is documented here. Overall, the SEJ exercise provided some benefit in calibrating the expected contamination rates for the purposes of developing the simulation study. It is not clear at this point whether the exercise was sufficiently valuable to justify its recommendation in further such exercises.

4.4 Future slippage survey design

We undertook a number of steps to develop a recommendation for a slippage survey sample design. Briefly, the process was:

1. workshops for discovery of pathway perceptions;
2. regulatory data analysis and modeling to understand pathway risks;
3. structured expert judgement to augment the data resources, and identify and estimate the known unknowns;

4. sample design; and
5. simulation study to assess different design options and sample sizes.

It's possible that the SEJ workshop was overly-detailed for the purposes to which it was put. It is noteworthy that the experts considered that there might be some differences between source countries worth mentioning, but the magnitudes did not seem operationally relevant. Future slippage survey exercises may be able to apply a streamlined version of this approach.

References

ACERA (2010). Process manual — elicitation tool. Technical Report Project 0611 report, Australian Centre of Excellence for Risk Analysis, The University of Melbourne, Melbourne. `http://cebra.unimelb.edu.au/data/assets/pdf_file/0010/1290466/0611-process-manual.pdf` (note that the Elicitation Tool to which the report refers is no longer available from the web site).

Adams-Hosking, C., McBride, M., Baxter, G., Burgman, M., Villiers, D., Kavanagh, R., Lawler, I., Lunney, D., Melzer, A., Menkhorst, P., Molsher, R., Moore, B., Phalen, D., Rhodes, J., Todd, C., Whisson, D., and McAlpine, C. (2016). Use of expert knowledge to elicit population trends for the koala (*Phascolarctos cinereus*). *Diversity and Distributions*, 22(3):249–262. DOI: 10.1111/ddi.12400.

Burgman, M., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L., and Twardy, C. (2011). Expert status and performance. *PLoS ONE*, 6:e22998.

Carey, J. and Burgman, M. (2008). Linguistic uncertainty in qualitative risk analysis and how to minimize it. *Annals of the New York Academy of Sciences*, 1128:13–17.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons, Inc., 3rd edition.

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Environmental Ethics and Science Policy Series. Oxford University Press.

Cooke, R. (2013). Expert judgement assessment: quantifying uncertainty on thin ice. *Nature Climate Change*, 3:311–312.

Dalkey, N. and Helmer, O. (1963). An experimental application of the Delphi Method to the use of experts. *Management Science*, 9:458–467.

Dodge, H. (1943). A sampling inspection plan for continuous production. *The Annals of Mathematical Statistics*, 14(3):264–279.

Epley, N. and Gilovich, T. (2006). The anchoring-and-adjustment heuristic. *Psychological Science*, 17:311–318.

FAO (2008). International Standards for Phytosanitary Measures: Methodologies for Sampling of Consignments. Technical Report ISPM 31, Secretariat of the IPPC.

Fleishman, E., Burgman, M., Runge, M., Schick, R., and Kraus, S. (2016). Expert elicitation of population-level effects of disturbance. In Popper, A. and Hawkins, A., editors, *The Effects of Noise on Aquatic Life II*, volume 875, pages 295–302. Advances in Experimental Medicine and Biology.

Galton, F. (1907). Vox populi. *Nature*, 75:450–451.

Gigerenzer, G. (2004). Dread risk, september 11, and fatal traffic accidents. *Psychological Science*, 15:286–287.

Gustafson, L., Antognoli, M., Lara Fica, M., Ibarra, R., Mancilla, J., Sandoval, O., Enriquez, R., Perez, A., Aguilar, D., Madrid, E., Bustos, P., Clement, A., Godoy, M., Johnson, C., and Remmenga, M. (2014). Risk factors perceived predictive of ISA spread in Chile: Applications to decision support. *Preventive Veterinary Medicine*, 117:276–285.

Hemming, V., Burgman, M. A., Hanea, A., McBride, M., and Wintle, B. (2017). A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution*, 9(1):169–180.

Janis, I. (1971). Groupthink. Psychology Today. 5: 43-76. In Shafritz, J., Ott, J., and Jang, Y., editors, *Reprinted in: Classics of Organization Theory.* 2015 Cengage Learning, Independence, KY., 8th edition.

Jarrad, F., Barrett, S., Murray, J., Parkes, J.and Stoklosa, R., Mengersen, K., and Whittle, P. (2011). Improved design method for biosecurity surveillance and early detection of non-indigenous rats. *New Zealand Journal of Ecology*, 35:132–144.

Krinitzsky, E. (1993). Earthquake probability in engineering - part 1: The use and misuse of expert opinion. *Engineering Geology*, 33:257–288.

Linstone, H. and Turoff, M. (1975). *The Delphi Method: Techniques and Applications.* Addison-Wesley Pub. Co., Reading, MA.

Maier, N. (1967). Assets and liabilities in group problem solving: the need for an integrative function. *Psychological Review*, 74:239–249.

Martin, T., Burgman, M., Fidler, F., Kuhnert, P., Low-Choy, S., McBride, M., and Mengersen, K. (2012). Eliciting expert knowledge in conservation science. *Conservation Biology*, 26:29–38.

McBride, M. F., Fidler, F., and Burgman, M. A. (2012). Evaluating the accuracy and calibration of expert predictions under uncertainty: predicting the outcomes of ecological research. *Diversity and Distributions*, 18(8):782–794.

Mukherjee, N., Hugé, J., Sutherland, W., McNeill, J., Van Opstal, M., Dahdouh-Guebas, F., and Koedam, N. (2015). The Delphi technique in ecology and biological conservation: applications and guidelines. *Methods in Ecology and Evolution*, 6:1097–1109.

Nisbett, R. and Wilson, T. (1977). Halo effect: evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35:250–256.

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities.* Wiley, London.

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, 29:261–265.

Plous, S. (1993). *The Psychology of Judgment and Decision Making.* McGraw-Hill, New York.

Robinson, A., Bell, J., Woolcott, B., and Perotti, E. (2012). DAFF Biosecurity: Plant-Product Pathways. Technical Report 1001J 1, Australian Centre of Excellence for Risk Analysis.

Soll, J. and Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30:299–314.

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., and Burgman, M. (2010). Reducing overconfidence in the interval judgments of experts. *Risk Analysis*, 30:512–523.

Surowiecki, J. (2005). *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* Doubleday, USA.

Sutherland, W. (2006). Predicting the ecological consequences of environmental change: a review of the methods. *Journal of Applied Ecology*, 43:599–616.

Sutherland, W. and Burgman, M. (2015). Use experts wisely. *Nature*, 526:317–318.

Teigen, K. and Jørgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology*, 19:455–475.

Thorndike, E. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4:25–29.

Tversky, A. and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology*, 5:207–232.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty — heuristics and biases. *Science*, 185:1124–1131.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211:453–458.

Wintle, B., Mascaro, S., Fidler, F., McBride, M., Burgman, M., Flander, L., Saw, G., Twardy, C., Lyon, A., and Manning, B. (2012). The intelligence game: assessing delphi groups and structured question formats. In *Proceedings of the 5th Australian Security and Intelligence Conference, Novotel Langley Hotel, Perth, Western Australia, 3rd-5th December, 2012.* `http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1025&context=asi`.

A Initiation Workshop Notes

Meeting Notes Initial Workshop, 1 pm 5/8/2020

1. Introductions, roles, and interests (AR)
   Participants

   - Michael Ormsby – Manager Plant Science and Risk Assessment
   - Andrew Robinson – CEBRA
   - Mervyn Alexander – Team Manager Border Clearance Services Metro
   - Chris Denny – Team Manager Border Clearance Services Freight & Mail
   - Deepak Kumar Gupta – Senior Business Analyst
   - Allan Burne – Senior Adviser Plant Science and Risk Assessment
   - Jenna Guest – Team Leader Intelligence and Targeting (ITT)
   - Sina Waghorn – Specialist Adviser Treatments & Inanimate Pathway (IHS)
   - Stu Rawnsley – Manager Biosecurity Clearance Services
   - Andrew Spelman – Manager National Programmes
   - Evelyn Pleydell – Team Manager Biosecurity Intelligence Team
   - Nikola Merrilees – Team Manager Assurance
   - Karen Pugh – Manager Horticulture Imports (IHS)
   - Francie Olliver – Manager Pathway Assessment (IHS)

   Apologies: Fiona Stewart (ITT)

2. Discuss pathways with material slippage results (Group)
   Purpose to find confidence in knowledge around slippage levels
   Definitions:

   - Target Slippage: Management slip (false negative)
   - Inspection slippage: Inspection slip
   - Conveyance slippage: vessel/container slip

3. Precedents for and purpose of slippage audit (NB: measure / verification!) (AR)
   Pathways currently under slippage surveys:

   - Containers — most (99%) of items are inspected or certified and there is a 2% audit of container and cargo compliance at transitional facilities with AP's present
   - Express — audit to test assumptions (Express freight = fast freight)
   - Cruise passengers — tests accreditation schemes (audit)
   - Passengers — minister set performance target (KPI) of 98.5% compliance
   - Mail — minister set performance target (KPI) of 99% compliance

Containers: slippage is being measured because it is a requirement of the IHS and because verification of Quarantine Declaration accuracy are undertaken by a third party (Accredited Persons; AP). There is also a lack of confidence in certification of cleanliness.

Latest figures from container audits — 7% failures for empties and 13–14% for loaded. Debate around what makes a risk (is the thing a risk). If not compliant then exporter gets escalated to next 5 inspected. As audit is done at APs, this is not independent observation (it is not clear what happens with APs when no inspector watching). Four-sided inspections for majority. APs only report on non-compliance. Random selection by system (Cusmod). APs report 3–4%, MPI finds 13–14%; difference could be in definitions.

Any implied or explicit triggers that would change interventions? Detection of BMSB (50 detections, 11 live high-risk pest) triggered change in actions (treatment). Sliding scale of intervention for empty container pathway (poor history of shipping line for std MT containers, port of origin (with approved cleaning system in place) for sea containers covered by the Sea Container Hygiene System / GAS).

Express Freight: Baseline survey — moved to ongoing random intervention schedule (dogs and officer checks). No formal measures at moment. Measures could be a compliance survey system. 3000 articles from the ICRs.

Measures apply to risk goods on pathway (not to pathway), so if missed there would be no intervention. Audit of pathway after system is difficult (currently at beginning of pathway on arrival). So, not a slippage survey (intervention happens randomly based on Inward Cargo Reports or ECRs (electronic cargo information) (for low value goods). Approximately 100 items per month are issued with a retrospective BACC for clearance because they have arrived at a CF or a TF without clearance and the facility notifies us.

Approximately 60 percent of those come through express freight and the remaining 40% via mail. We also have an approved permit scheme with some CFs and TFs that effectively authorises the TFs and CFs to keep a record of the items they receive over a month with no BACC and apply for a single BACC monthly to cover them. This scheme is audited by ITT to ensure compliance.

High Risk or High Value Express freight must be declared. Targeting/inspection is based on Manifest Rather than Declaration.

Thresholds: Actions if not meeting target. Also, likely to change if high risk pest(s) or high-risk goods detected (in volume).

Cruise Passengers: Thresholds — 98.5% compliance target set by Cabinet, but likely to change if high risk pest(s) or high-risk goods detected (in volume). Cruise slippage survey designed to test effectiveness of accreditation of cruise ships — so its function is verification rather than intervention.

What is indicator of failure? Information on shipboard management — e.g. where are stores sourced etc. We might send a staff member offshore to audit.

Every cruise ship has an offshore accreditation audit completed offshore prior to their first arrival of the season. Over the course of the season there is also a 'performance verification' (PV) schedule where vessels are audited at various ports to ensure they are still meeting the requirements of the accreditation scheme. There is also a gangway intervention schedule which is based on various factors including history, 1st port of arrival...? Escalation of interventions applies dependent on the result of both gangway interventions and the PV survey. Also each vessel individual vessel has a threshold for

seizures based on pax numbers which will trigger increased interventions on that vessel and can lead to potential cancellation of their accreditation.

Op top of this there is the std intel/targeting work that is done assessing compliance with the Craft Risk Management Stds.

Passengers: Thresholds — 98.5% compliance target set by Cabinet, but likely to change if high risk pest(s) or high-risk goods detected (in volume). We also look at comms from cruise line to passenger about what they can or cannot bring ashore. Mail: Thresholds — 99% compliance target set by Cabinet, but likely to change if high-risk pest(s) or high-risk goods detected (in volume).

4. What do we know about express freight? Data? Anecdotes? Risks? (Group)
   Data from slippage surveys (IMP survey — 1 in 10 random sampling of goods not being stopped — 2.5 years). Found very little therefore considered low risk. Shoes with 24 live BMSB came through pathway. What is found on mail pathway is often undeclared or misdeclared. In 2018 found seized parcels at the IMC were: declared (35%), not declared (34%), misdeclared (28%) or unknown (3%). There are also mail survey results and raw data going back to 2013 if required.

   IMC QOs record what terms are misdeclared on packages when risk goods are found so we can track declaration trends. Electronic data associated with goods, and measures for those.

   Difficult to get these results in appropriate form for analysis, would allow for better understanding of risk across system.

   A group of officers use a survey approach that is both random (manual selection of parcels off a belt using a random selection methodology) and targeted — dogs used as a detection tool, possibly x-ray in some sites if facilities allow.

5. Identify audit outcomes e.g., who needs to know and why (AR) Intel, risk, IHS, policy. Process improvement is purpose. To audit compliance.

6. The difference between counts (measure) and rates (verification) (AR) Pathway compliance rate = how well we are managing pathway risk.

   Counts = how many non-compliant passengers are passing through the system — shows how big the risk is. Discussion around whether we should report numbers rather than proportions — need to measure actual risk associated with non-compliance. Knowing what happens makes numerical data less confronting — numbers would be used to determine acceptable level of risk. Would slippage numbers be used to improve system function and efficiency? MPI have made significant operational changes based on the learnings from slippage numbers. Additionally new comms campaigns have been developed as a direct result of survey findings, and multiple changes to stds have been made and CTODs specific to accompanied goods were created.

   Rates (%s) vs. actual numbers — numbers better reflect actual risk as not volume dependent for context. Could also look at level or risk of each event. Risk model on numbers/propagule pressure (IBRAM).

7. Expert Elicitation: proscribe trigger slippage levels for express freight (Group) Acceptable level of Slippage. Poor or incorrect information is being given to target evaluations. Merv commented that Target slippage is the biggest problem — system won't allow effective targeting of low value express freight. About 6% slippage where manifest declaration does not

match contents (over numerous companies; quarterly reporting, freight companies are responsible). The 6% figure came out of the express freight risk determination survey that was completed Dec 2019/Feb 2020. We have included data collection around manifest accuracy into the BAU random intervention schedule — Chris Denny's area. Our expectation is that we report quarterly back to express freight companies — they are accountable for resolving as much of the 6% as possible.

Parcel declaration feeds into manifest declaration; 40% manifest declaration didn't match parcel declaration.

34% of parcels had no parcel declaration. However only 2 out of 3000 inspected presented a biosecurity risk. Mis-declared items (e.g., shoes declared as clothing) were generally clean. Once deselections removed this number came back to about 2300(?)

Customs is also interested in mis-declaration. Question: what level of mis-declaration is acceptable?

It was suggested the level of acceptance should vary among pathways considering the volume and value [impact] of biosecurity threat $\leq 5\%$ proposed as a level, but it was noted that the acceptable level of slippage/misdeclaration posed by a low volume pathway was not the same as on a high-volume pathway. Express freight tripled in volume over 3 years.

A deviation or increase of 1% was considered to be cause for concern and it was noted the rate of slippage may not be representative of the rate of non-compliance.

All types:
Specific types: incorrect information (target slippage) — 6% now, but not known what problem looks like ("6% looks like low")("Stu thinks 5% is threshold")("Customs are also not happy with these numbers (threshold unknown)"). Random intervention data looks at what is causing the problem. Parcel declaration vs. manifest declaration (e.g. either wrong or mismatch). 1 in 3,000 was mis-declared and potentially a risk (out of 167).

Volume of pathway tripled over 3 years so is rate appropriate.

If high risk pest(s) or high-risk goods detected (in volume), then response?

Change intervention over Xmas as goods labelled "gift" is not informative so all such labelled are x-rayed.

Significant deviation beyond 1% (99.96% of items are non-risk) — e.g. "gifts" is a deviation. 1% slippage of risk goods (could be less e.g. 0.2%). Would need to be more nuanced as the nature of the risk good reflects tolerance.

Random seeds on mail pathway is normal, but recently political.

Random note saying "18% contamination rates for empty containers and 22% for loaded containers — The IHS stipulates 'Clean and Free from Regulated Pests' rather than having a specific Regulated Pest" There is an escalation from Level 1 audit to Level 2 following non-compliance. Five consecutive compliant consignments are required to move back to level 1.

There was some discussion around the presence of an auditing quarantine officer (QO) on how thorough AP inspections are. It was suggested that results of an audit may not accurately reflect rates of slippage. It was also noted that Aps and QOs report different rates of slippage (APs report 4% and QOs 13–14%).

An AP is only obliged to report if a container if a container is not clean. High Risk containers are inspected by QOs rather than APs. "High Risk" containers are containers with either

high risk of contamination by such pests because of origin/history with a particular pest (e.g. BMSB or Giant African Land Snail, GALS) or from an exporter with high levels of non-compliance. Non-compliant container declarations trigger a QO supervised devann. Also country of origin will pre determine if they are high risk as per the identified list of high risk countries but this is specific for MTs.

Detections (e.g., BMSB associated with Italian Sea Containers) trigger a targeted response and increase inspection frequency — Continuous Sampling Plan (CSP1)

Fewer than ten BMSB [mass] detection events triggered additional measures (Compulsory Methyl Bromide or Heat Treatment).

8. Break: 10 minutes
9. Discussion: counterfactuals (establish losses) (Group)
10. Discussion: sources of variability (time / space) (Group)
11. Expert Elicitation: estimate actual slippage levels for express freight (Group)
12. System design (Group)
    (a) identify frame,
    (b) sample selection,
    (c) sample inspection,
    (d) data collection,
    (e) reporting,
    (f) limitations
13. Debrief, Questions, Next Steps Operational lead: Nikola Merrilees; Science lead is Mike Ormsby.

B SEJ Workshop Participant notes

The following summarises the meeting held on August 4, 2021. Section B1 considers the questions mentions in Section 2 of this report. Section B2 gives the minutes of the workshop and Section B3 reports some comments made after the workshop.

B.1 Round 1 notes

B.1.1 Question 1: upper risk threshold

"1. Imagine 1000 random UPE consignments coming from a range of countries to NZ. What is the lowest number of consignments that require intervention beyond standard inspection that would convince you that a different set of measures or policies are required to properly manage the biosecurity risk associated with UPE? That is, what is the lowest of the numbers that would say to you: this pathway is unacceptably risky and we have to do something!"

- Using a volume of between 10 and 20%, which is in line with auditing checks on sea containers. This also takes into account the severity of the risk imposed.
- Air cargo is less likely to be of risk than sea cargo.
- To be honest I believe the % we currently inspect is fairly correct and in fact could be decreased somewhat. Current rate of inspection is about 35% and has a hit rate of actionable risk of only 6%.
- If I make an assumption that only 3rd (0.333) of total UPE consignments (1000/3) need intervention beyond standard inspection at random; I could say if I had 2/3rd of the original 1/3rd numbers that could be conforming post inspection then I'll be happy. And the fact that I would only have 1/3rd (0.11) of numbers needing intervention beyond std. inspection hence giving me around 10% to 11% of original number. That is less than 90% confidence that we got all the risks.

B.1.2 Question 2: lower risk threshold

"2. Imagine 1000 random UPE consignments coming from a range of countries to NZ. What is the highest number of consignments that require intervention beyond standard inspection that would comfort you that a lesser set of measures or policies are sufficient to manage the biosecurity risk associated with UPE? That is, what is the highest number that would say to you: this pathway is surprisingly safe! Let's do less here and focus our attention elsewhere!"

- This amount I believe we could rely on AP's and importers to alert MPI about and be reasonably comfortable with.
- I believe there is still a fair volume of contaminant found that might not be reported as much as it should be because it is easier to manage at the time of inspection — therefore it seems imprudent to drop our intervention level too low — but maybe requires regular adjustment.
- I think I'm more risk averse. I feel like others numbers will be higher.
- If I had more than 1/10th of the total 1/3rd of the UPE consignments down for standard inspection fail. That is more than 97% confident that we got all the risk.

B.1.3 Question 3: low-risk goods

"Imagine 1000 random consignments of UPE from all countries. Their documentation corresponds to low-risk (that is, risk score ≤ 5 and no must-inspect items). How many of those 1000 random consignments will have actionable risk material?"

- 2 to 3 out of 10 low risk consignments are likely to have risk materials from my past experience; but risk might vary from country to country. But we are not looking for anything when we are not looking for anything!
- High — there could risk items in there that just don't meet the freshold to inspect but could still present risk .
  Low — The majority of a person's household items are not of biosecurity risk (by comparison to risk items).
  Best — In the middle
- High — documentation could indicate that care has not been put in.
  Low — just bad documentation that is not indicative of anything
- Assume my high number as being a similar interception level as cargo we do send for inspection. Assume my Low number as being approx 2% contam level. This is based on the fact that there are large volumes of very low risk goods in these shipments, e.g. boxes of clothes, books, kitchen items, etc.
- Christmas decorations should be mandatory from all countries and or more countries like Canada added to the list, if the MPI supplementary form were mandatory we might not miss as much, as the form prompts importers to think more about what are risk goods and what is in their package. Air freight has fewer risk items for UPE's than sea.
- We are relying on importers' declarations. Experience from the airport environment has shown that assistance is needed to prompt correct answers. Also importers fill in forms retrospectively so the memory or knowledge of the risk items may not be correct.
- I have increased a few countries due to their (in my opinion) being more likely to be full households rather than selective items. Full households are collected and/or packed from contracters who enter the premises and pack and load everything. The chance that goods are not clean or items were sent that weren't meant to be is higher.

B.1.4 Question 4: high-risk goods

"Imagine 1000 random consignments of UPE from all countries. Their documentation corresponds to high-risk (that is, risk score > 5 or at least one must-inspect item). How many of those 1000 random consignments will have actionable risk material?"

- The number corresponds mainly from the bigger volume countries, which are more likely to be Full households with more items and more actionable risk material.
- Past experience has shown the actionable items may be minimal but still necessary.
- These countries have increased numbers due to the importers are more likely to camp, hike and garden which are higher risk.

- Pet bedding possibly should be mandatory, it would be good to know breakdown on what the actionable risk material is, CITES stuff is also found a lot in UPEs, Pacific islands UPE data might not be accurate as we enter a lot under plant products or food lines rather than UPEs as they usually are TAPA mats or frozen produce, handicrafts.

- For these calculations I literally used the highest number of contamination (49%) and the lowest level of contamination (1%). My Best was remaining as the current average (6%).

- High — documentation could indicate that care has not been put in.
  Low — just bad documentation that is not indicative of anything

- High — As part of the risk scoring system select items have been identified as being a risk commodity or likely to be harbour risk. Based on previous finds there's a higher likelihood of these items actually presenting actionable risk once inspected.
  Low — Risk goods that require action under the scoring system only make a small % of a person's household items which are not of interest. Plus some of those risk items may not be shipped back to NZ or they don't have eg. UK has high volumes of UPEs but in big populated areas like London people may not have risk items like Lawn mowers/BBQs due to apartment living.
  Best — Based on the data in the following sheet on what actionable consingments have been found over the previous couple of years.

- If consignments are flagged as high risk we would be actively looking for things even if they are minimal risks. UK has relatively high risk factor with used goods. SA would again be because of relative nature of used goods being brought in. Again UPEs from India majorly will have plant products that are used.

B.2 Workshop Minutes

B.2.1 Thresholds

Inspecting certain amount; still not finding much risk. Hence, confident that we inspect enough. If 35% found 35% then that would be different. But we find less.

Anything over 2% should be inspected — what about less than 2%? What is the 2%?

Thresholds depend on risk aversion. Also about having information about what is in the pathway - even having moved house. This is really about the biosecurity risk. Inspecting for years; not evidence of higher biosecurity risk. It's about the actual biosecurity risk as opposed to potential for risk - outdoor furniture and garden tools. How much actionable BR? 60% that is examined is non risk 30% close to risk. A lot of that comes to not being able to talk with the importer, cf airport situation - interview changes apparent risk. Documentation only. What things are you moving with? Things that are dear to you. Mostly well kept. Also folks moving for job, and if subsidised they'll bring everything. Stratified risk!?

B.2.2 HR pathway

Depends on meaning of docs. Could be no care? Relocation reason. Having screened hundreds of manifests — boxes of books, clothes, pans. Definitely standard things. Often OS moving company filling in the documentation, very commonly done. Moving company does final packing list. Importer often fills out docs a month before move, or remotely. Categorising things: eg sports equipment vs HR wooden items. Country of origin plays into it. Eg BMSB season — clothing from

USA could be a risk item. Rural / urban settings. List doesn't change much based on country of origin. Unlikely to submit all cartons of clothing from USA. Where has the clothing been stored?

Country-specific: AUS, UK, SA, USA tends to be more full household movement. Pacific Islands: relying on evidence on last page — those countries have low volumes of importing but high rates of finding risk. Lots of tapa mats and those are risky. High proportion of goods that are religious / cultural. India: could be lower; consignments are small. UPE risk is food items through airport. All food goes for inspection but comes through express freight. Don't use a UPE form. Pretty badly contaminated.

B.2.3 LR pathway

Note earlier comments. Manifests with boxes of books, clothes, DVDs, kitchen ware. Rarely have high potential of being contaminated. How is stuff declared. Folks might decide not to import things they know to be of biosecurity risk. Even one risk item can still be risky. Trying to believe documentation — honest and accurate. Who made the declaration? Removal company. Too many people involved — could confuse the declaration, eg details, condition, might come from separate sources. Cleanliness is subjective; interpretation.

Pacific Islands: can't procure much; limited amount of goods. Majority locally procured. Not as much at risk of biosecurity incursions. India, Philippines: part of reasoning is the countries get more scrutinised because of features of the goods seen on regular basis.

Quite some slippage in UPE. Data highlights it. Personal impressions suggests HR. Might be multiple processors for documentation. Lack of real-time information; no interview. Future: target evaluators should do some out and about with UPE — meaning, be involved with HH delivery. Eyes opened to risk items let through. Find the items when they are unwrapped. Experience with what's delivered. However, evidence needs to be from what is being found as well., Data entry might need to be stepped up. An AP devans a container but doesn't deliver full household. Hard to evaluate true risk.

B.3 Follow-up notes

These notes were provided after the workshop.

- Risk commodity lines being created instead of UPE lines (i.e. plant products lines or animal products being created, even though UPBD is presented and line should technically be UPEs) — this is why I put Fiji, Tonga, Samoa etc as having higher actionable items even though perhaps the data does not show this — usually these UPEs are small air freight consignments that only have a few items and are accompanied by phyto- or zoo-sanitary certifications (Kava powder, tapa mats, cocoa, marine fish) i.e., I believe their documentation to be more trustworthy than an importer's packing list for an FCL.

- I believe that the data for certain products may also have changed in the past year — in December 2019, the Import Health Standard for Dried and Preserved Plant Material was amended, and unaccompanied Pacific style mats no longer required inspection after presenting an original phytosanitary cert — I am not privy to the effects this has had on interventions (i.e. the contamination/ARM remains the same, but inspections are not being carried out to find the contamination/ARM).

- Importers having no understanding or knowledge of biosecurity and therefore not declaring their risk goods accurately. Not every country prioritises biosecurity to the same extent in my experience and importers do not understand the risks so will not be as particular about their declarations.

- Personal experience in speaking to importers has made me very risk adverse for UPE consignments — often they cannot remember exactly what was packed OR they do not accurately declare (purposefully or not). A recent example — spoke to importer on phone as they were trying to send in a packing list. I took the opportunity to ask about the nature of items and if plant material was present (herbal tea (green) and snacks were declared only). The tea turned out to be full of hemp seeds and other dried plant materials were present that had not been declared. I think I am very risk adverse, especially for risk assessment procedures that are based on purely documentation provided by importers who may have no understanding or appreciation for biosecurity.

- Consignments from China are often fumigated offshore with methyl bromide, which may lead to lower ARM being present/interventions.

- ARM from countries like USA/GB may have higher recorded interventions, as (I am assuming here!) the products we keep an eye out for require devitalisation treatments by approved treatment provider (such as pine cones, plant materials on Christmas decorations). This will be recorded more accurately in QuanCargo with a separate authority, unlike an officer removing leaf material and placing it into a quarantine bin onsite.

C SEJ Briefing notes

This document is copied with only modest alteration from CEBRA report 1505A Output # 1,[5] the germane text of which was authored by Dr. Jan Carey of the Australian Centre of Excellence for Risk Analysis (ACERA). Any inconsistencies are the responsibility of the first author of the current document.

Ideally, decision-making should be informed by high quality data. However, empirical data are often sparse or lacking in fields such as conservation biology, environmental management (Martin et al., 2012) and biosecurity (Burgman et al., 2011). In such cases, managers and other decision-makers have little option but to rely on the opinion of experts (Sutherland, 2006). Increasing use of expert judgement (Martin et al., 2012) has stimulated the development of elicitation methods that aim to address recognised problems and thus improve the quality of judgement obtained from experts (Sutherland and Burgman, 2015). Structured expert elicitation is increasingly recognised as a valid approach in data-poor settings (Cooke, 2013).

C.1 On expert judgement

The use of expert judgement is difficult. Individuals are known to be subject to a number of psychological frailties and cognitive biases, including the following:

- Framing — The way a question is presented or framed can influence the response received (Tversky and Kahneman, 1981; Plous, 1993).
- Availability — Judgements of probability may be influenced by the ease with which an event is recalled. Events may be judged to have a higher probability of occurrence than is really the case if they are recent, invoke strong emotions or have been widely reported (Tversky and Kahneman, 1973; Plous, 1993; Gigerenzer, 2004).
- Anchoring and adjustment — The tendency to base quantitative estimates on values that been previously suggested or estimated (Tversky and Kahneman, 1974) coupled with an inability to adjust sufficiently far from the anchor (Epley and Gilovich, 2006).
- Overconfidence — When experts have unwarranted confidence in their own judgments (Oskamp, 1965).

Using a group of experts instead of just one individual draws on a wider range of experience than could be achieved with just a single individual and avoids judgements that reflect only the cognitive biases of a single individual. An early example of what Surowiecki (2005) has labelled "the wisdom of crowds" dates back to an agricultural show in the early 1900s when 787 entries in a competition to judge the dressed weight of an ox were summarised by Galton (1907). The median of the entries was less than 1% from the true weight of 1198 lb.

Of course, group assessments are also subject to their own suite of biases, including:

- Group think — Judgements are unduly affected by a desire for agreement within the group (Janis, 1971).
- Dominance — Individual experts are unduly influenced by the views of a senior or dominant group member (Maier, 1967).

[5] Ornamental Fish Import Reform — Health Monitoring Program

- Halo effects — Perceptions of the opinion of one expert are influenced by the perception of attributes of that expert unrelated to the subject under consideration (Thorndike, 1920; Nisbett and Wilson, 1977).

The use of structured expert elicitation can minimize the impacts of these biases.

C.2 Structured expert elicitation

If expert judgement is to be used to inform decision-making, then clearly it is desirable that the judgement be of the best possible quality. Quality in this context has been defined in terms of calibration between judgements and reality (O'Hagan et al., 2006) and the degree of precision and confidence associated with an estimate (Cooke, 1991). Some key elements of structured expert elicitation approaches are described below:

- The Delphi technique was designed to obtain a "most reliable consensus of opinion of a group of experts" (Dalkey and Helmer, 1963). It was developed at the RAND Corporation in 1950s as a predictive tool for military purposes. More recently, it has been widely used for expert elicitation in some disciplines such as medicine and social policy, but is used rather less in ecology and conservation science (Mukherjee et al., 2015).

  A standard Delphi assessment involves two or more rounds of estimates from a group of experts. Between rounds, an anonymous summary of judgements and the reasoning behind the judgements are provided to participants, who are then able to revise their estimates if they wish (Linstone and Turoff, 1975). The private nature of judgements (e.g., completion of a questionnaire) aims to reduce dominance and halo effects; individuals are not directly influenced by others in the group they might perceive as more expert.

- Questions asked of experts can be structured to minimise the experts' overconfidence. Soll and Klayman (2004) proposed a three-step format for estimates of quantities (lower, higher, mid-point for researcher-defined level of confidence) which reduced overconfidence compared to that associated with eliciting ranges (intervals) or point estimates.

  Speirs-Bridge et al. (2010) extended the format to four steps with the addition of a measure of confidence assigned by the expert. Providing lower and upper bounds separately (Steps 1 and 2) encourages experts to consider different lines of evidence before moving on to make their best estimate (Step 3). Because people are generally better at evaluating intervals than producing them (Teigen and Jørgensen, 2005), the assigning of their own level of confidence to their interval (Step 4) offers experts an opportunity to review that interval. Calculation of 'derived' intervals at a common level of confidence (generally 80%) is then necessary for the purposes of comparing and combining the intervals of individual experts (Speirs-Bridge et al., 2010).

- Not all group judgements will be accurate and well calibrated with the truth (Krinitzsky, 1993). Surowiecki (2005) identified four conditions that characterise 'wise crowds' able to generate good group judgements, namely: (i) diversity of opinion, (ii) independence, (iii) decentralisation (individuals draw on their own local knowledge) and (iv) aggregation (having a suitable means to generate a group judgement from multiple individual estimates). Where there are calibration questions to provide a measure of the performance of individual experts, the responses of experts might then be weighted to improve the quality of group judgements (Cooke, 1991).

C.3 IDEA: Modified Delphi-style elicitation

Researchers associated with the Centre of Excellence for Biosecurity Risk Analysis (CEBRA; formerly the Australian Centre of Excellence for Risk Analysis, ACERA) developed a Delphi-style elicitation approach of private judgement / group discussion / private judgement that allowed for the sharing of evidence while also addressing some cognitive biases known to have a negative effect on group judgements (ACERA, 2010). This protocol, named IDEA, is described in the following section.

The discussion phase is more direct than with the standard Delphi method. While summarised data are still presented free of personal identifiers, participants are able to communicate with one another, for example, via group emails or in a workshop setting. This provides an enhanced opportunity to test the credibility of particular estimates and the logic behind them.

CEBRA's Delphi-style elicitation process also differs from the standard Delphi in not requiring consensus among the experts. Emphasis is placed on minimising language-based uncertainty (Carey and Burgman, 2008) and identifying any assumptions underlying the questions asked of the experts. Genuine differences of opinion that remain after discussion are acknowledged and incorporated in the final outcome, a key aim being to quantify the extent of uncertainty.

Various aspects of CEBRA's approach to structured expert elicitation have been tested experimentally, with the following results:

- In experimental studies where the truth was known to researchers but not to study subjects, group judgements were generally found to perform as well as or outperform even the best-regarded individuals in terms of accuracy and calibration with the truth. Domains tested were the prediction of geopolitical events (Wintle et al., 2012) and questions of human health, biosecurity or ecology (Burgman et al., 2011).

- The accuracy of group averages improved substantially after the discussion stage in six separate workshops dealing with questions of human health, biosecurity or ecology (Burgman et al., 2011).

- No consistent relationships were found between elicitation performance and years of experience, publication record, self-assessment of expertise or any other demographic measure of expert status in questions of public health (Burgman et al., 2011), biosecurity (Burgman et al., 2011; McBride et al., 2012), ecology (Burgman et al., 2011; McBride et al., 2012), or geopolitical events (Wintle et al., 2012). This was in contrast to the initial expectations of the participating experts in the studies of Burgman et al. (2011).

- The four-point elicitation method of Speirs-Bridge et al. (2010) demonstrably reduced expert overconfidence when tested in the domains of ecology (Speirs-Bridge et al., 2010; McBride et al., 2012), and epidemiology (Speirs-Bridge et al., 2010).

C.4 Description of the IDEA protocol

Briefly, the IDEA protocol involves four stages: Investigate, Discuss, Estimate, and Aggregate, as follows.

Investigate In this first phase, experts work individually using whatever resources are available to them to formulate point and interval (low—high) estimates of the unknown quantities. This step is akin to crowd-sourcing the estimation process. It should be preceded by a conference

call or meeting with all experts to identify ambiguities, agree on meanings, and affirm the elicitation approach and schedule.

Expert-elicited estimates are then anonymised and summarised by the facilitation team, to be used as spurs for discussion in phase 2.

Discuss The second phase is best carried out as a face to face workshop. Summaries of the anonymised phase-1 responses are presented to the group, with invitations to discuss the distribution of results — why might there be disagreement between experts, what might be reasons for particularly low or high values, etc.

Estimate The experts are then invited to privately update their first-phase point and interval estimates. They may choose not to do so. They are not obliged to disclose either way.

Aggregate The estimates from phase 2 are then aggregated within expert (to the extent that it makes sense to do so) and then across expert. For example, if the outcome of the exercise is to produce estimates of parameters for a structural model, for the purposes of making predictions, then a suite of predictions should be made for each expert, which should then be aggregated.

Hemming et al. (2017) provides a comprehensive description of the process.

C.5 Applications of expert elicitation

The use of quantitative data generated by expert elicitation is not without controversy. There are concerns that such data may be "biased, poorly calibrated, or self-serving" (Martin et al., 2012). Sutherland and Burgman (2015) emphasised that the quality of expert-generated data should be considered before using them in subsequent analyses or to inform decision-making. These authors go on to point out some of the strategies that can be employed to improve the quality of expert judgements, including the use of groups and a structured approach to the elicitation.

Some recent applications of quantitative data obtained by structured expert elicitation are outlined below:

- Expert opinion contributed to the development of a surveillance system for black rats on an island of high conservation value (Jarrad et al., 2011). The experts estimated parameters where data were lacking for the model used to optimise surveillance.
- Experts in fish health were used to establish likelihood ratios for risk factors that might influence the occurrence of infectious salmon anemia in Chilean salmon farms (Gustafson et al., 2014).
- About 20 experts of diverse backgrounds and affiliations participated in an elicitation exercise on the effects of collisions on North Atlantic right whales (Fleishman et al., 2016). The experts first contributed to the development of a conceptual model of collision effects, then provided estimates of the principal parameters needed to populate the model.
- A panel of 15 active koala researchers participated in a structured expert elicitation to estimate the total koala population of eastern and south-eastern Australia and changes in populations within different bioregions (Adams-Hosking et al., 2016). The entire elicitation process took six months, including a four-day workshop for the discussion phase.

D Technical details

From each expert, we elicited a best guess, as well as an lower and upper limit for the true contamination rate for each stratum, and a threshold above which risk management should change. (We also elicited a threshold below which risk management should become less intensive but did not pursue this angle further). For each stratum, we then converted their lower limit, best guess, and upper limit to a distribution best described as co-joined right-angle triangles, and simulated 10000 draws from the distribution (Figure 7)[6].

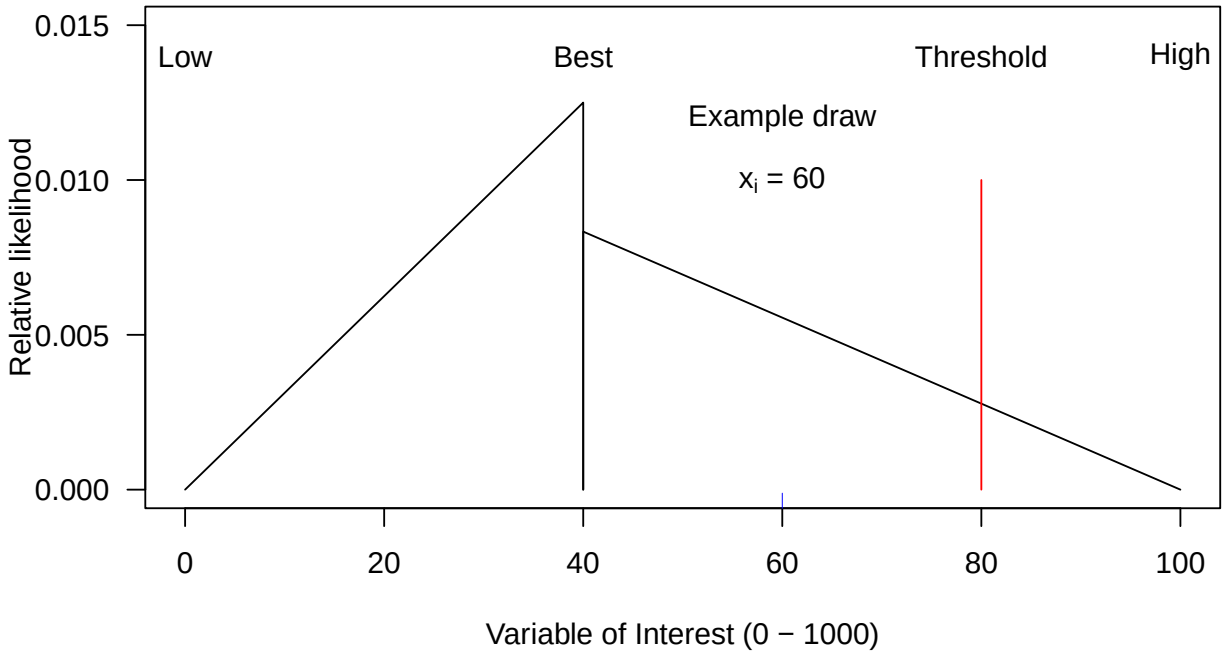

Figure 7: Example of elicited distribution. The lower right-triangle is delimited by "Low" and "Best", and the upper by "Best" and "High". Each has weight 0.5. The upper threshold is at 80 (red line), and an example random draw is shown at 60 (blue tick).

Their stratum-level draws were then weighted and summed to get an overall pathway-level estimated contamination distribution, say X_i $(i = 1 \ldots 6)$ for expert i, using the known stratum weights (approximately 2/3 for low-risk and 1/3 for high risk). These values were subsequently used to guide the following simulation study.

[6]This may seem like a counter-intuitive choice for the challenge of converting three numbers into a distribution, but extensive experimentation with other approaches in earlier work, including formal triangle distributions, beta-pert, and so on, suggests that this simple algorithm does the least violence to the outcome.

E Simulation study

E.1 Preliminary work

In order to specify details such as sample sizes, we first need to specify the way in which the data will be analysed when it is collected. That will enable us to assess the effects of various allocations of effort to the outcome.

Define the following pathway-level quantities as:

N is the total number of consignments in the pathway
y is the total number of contaminated consignments in the pathway
N_l is the number of low-risk consignments (approx. 2/3, none of which are inspected in BAU)
y_l is the number of contaminated consignments among the low-risk consignments,
N_h is the number of high-risk consignments (approx. 1/3, all of which are inspected in BAU)
y_h is the number of contaminated consignments among the high-risk consignments, and
z_h is the number of contaminated consignments detected among the high-risk consignments during BAU intervention;

and $y = y_l + y_h$ and $N = N_l + N_h$. Our goal is to estimate the following quantities, in the following order of decreasing importance:

1. the overall contamination rate of the pathway (y/N),
2. the actual relative risk (expressed as an odds ratio) of the high-risk consignments compared with the low-risk consignments ($\frac{y_h N_l}{y_l N_h}$, targeting), and
3. the slippage from the inspection ($y_h - z_h$).

We assume that contamination outcomes are independent and identically distributed, and that

$$y_h \sim \text{Binomial}(N_h, p_h) \text{ and } y_l \sim \text{Binomial}(N_l, p_l) \tag{2}$$

For the purposes of the design power, we also assume that the BAU inspection finds *all* of the contamination if it finds *any* contamination — that is, there are no consignments for which (i) the BAU inspection is performed, (ii) contamination is detected, (iii) the slippage survey is performed, and (iv) *more* contamination is detected. Accommodating this more complex outcome is possible but we decided it was out of scope for the current exercise, which is to design the survey. Note that such an analysis can be performed under the proposed design, so our assumption does not preclude it.

Denote the proposed slippage sample size from the high-risk stratum as n_h, and the slippage sample size from the low-risk stratum as n_l. The samples are simple random samples, so we can then denote the number of interceptions in the low-risk and high-risk strata as x_l and x_h respectively, and assume that slippage from the slippage samples is negligible.[7] Then, we can assume that

$$x_l \sim \text{Binomial}(n_l, p_l) \text{ and } x_h \sim \text{Binomial}(n_h, p_h). \tag{3}$$

Estimating p_l is straightforward; we can simply take $\hat{p}_l = x_l/n_l$ as a consequence of the specified survey design. Estimating p_h is tricky because there are two sources of information about it, namely the BAU inspection and the slippage inspection.

At this point several approaches are possible for estimating the high-risk rate, as follows.

[7] This assumption is reasonable for the DAWE-style inspections that are proposed for the slippage survey.

E.1.1 Ratio estimator

The ratio estimator estimates the high-risk contamination rate using the ratio of the slippage and the BAU results, as follows:

$$\hat{p}_h = \frac{z_h}{N_h} \frac{x_h}{z_h^*} \tag{4}$$

where z_h^* is the number of BAU detections in the slippage sample (the quantity estimated by $\frac{z_h^*}{x_h}$ is also called the inspection effectiveness or sensitivity).

E.1.2 Slippage only estimator

The unconditional estimator ignores the BAU inspection and estimates p_h the same way as p_l above, namely $\hat{p}_h = x_h/n_h$.

E.1.3 Conditional estimator

The conditional estimator uses the information in the BAU inspection as follows: stratify the high-risk stratum into two sub-strata based on the outcome of the BAU inspection, detections and non-detections. Use the slippage survey in the non-detection sub-stratum to estimate the probability of a false negative, and then estimate the number of false negatives in the non-detection sub-stratum.

$$\tilde{p}_h = \frac{1}{N_h} \left(z_h + (N_h - z_h)\hat{\rho} \right) \tag{5}$$

where $\hat{\rho}$ is formed as follows: considering only the BAU negatives that were in the slippage sample, compute the proportion that were positive in slippage inspection.

That is, the estimated number of contaminated consignments is the sum of the contaminated consignments detected by BAU and the estimated number of contaminated consignments missed by BAU, which is estimated as the product of the number of consignments that were not detected as contaminated and the probability of a false negative in BAU.

E.2 Targeting efficiency estimate

We advocate use of the odds ratio to estimate the targeting efficiency. Large-sample theory shows that the following holds approximately under specific conditions:

$$\log OR \sim \text{Normal}\left(\log OR, \frac{1}{n_h p_h} + \frac{1}{n_h(1-p_h)} + \frac{1}{n_l p_l} + \frac{1}{n_l(1-p_l)} \right) \tag{6}$$

This equation is derived for conditions that are not well met in the current setup, but the simulation study (see below) suggests that the interval estimates are very conservative regardless.

E.3 Examples of HR calculation

This section provides a simple example of the three estimators to clarify the differences. Table 3 sets out an example of data for the high-risk stratum.

Note that there are 10 units in the population, 4 hits arising from the BAU inspection, 5 units in the slippage survey, of which 3 are detections, and of which 2 were also hit in BAU. The slippage survey contains 3 units that were not BAU hits, and one of these was a slippage hit (hence the probability of a BAU false negative is estimated as $1/3 = 0.33$). We assume that a hit in BAU will always be counted as a hit in slippage.

Table 3: Worked example for candidate estimators for high-risk stratum contamination weight. BAU has X where the usual inspection detected contamination, Slip. Insp. shows X where the unit was included in the slippage inspection, and Slip. Int. shows X where the slippage inspection detected contamination.

Unit	1	2	3	4	5	6	7	8	9	10
BAU	X		X				X		X	
Slip. Insp.	X	X	X	X	X					
Slip. Int.	X		X	X						

Using these data, the estimators would be computed as follows.

1. From the BAU inspection only the usual estimate would be $4/10 = 0.4$
2. From the slippage inspection only the estimate would be $3/5 = 0.6$
3. The ratio estimator multiplies the BAU interception rate (0.4) by the ratio of the slippage interceptions to the BAU interceptions within the slippage sample ($3/2 = 1.5$) to get $0.4 \times 1.5 = 0.6$.
4. The conditional estimator stratifies the population by the BAU result and estimate the probability of a false negative using the slippage survey ($1/3 = 0.333$), resulting in $4 + 6 \times 0.33 = 0.6$.

E.4 Simulation design

We compared the three potential point estimators for accuracy and corresponding interval estimators for length and coverage using a simulation study. The assumed parameters for the reported run were as follows:

N, the total number of consignments in the pathway, was 20,000
0.33 of the consignments were high risk on average
p_l was 0.05
p_h was 0.10
the effectiveness of BAU inspection was assumed to be 0.7.

E.5 Outcome reporting

For each of the estimates we have the following considerations: interval coverage rate interval width, and overall estimate accuracy. Interval estimates are key because in a valid statistical sample, they provide information about the amount of information captured by the survey. A wide interval signals a less informative estimate than does a narrow interval. However, the interval estimates have to obey certain rules, which are in effect tested by a simulation study such as this one.

The interval coverage rate provides information about the legitimacy of the estimate. If the interval coverage rate is substantially less than nominal then interpretation of the estimate is fraught. If the interval coverage is too high then interpretation will be unnecessarily conservative, which is not necessarily too bad.

The interval width provides information about the relative efficiency of various estimates, or put another way, the interval width reports how well information is extracted from the data. A narrower width is preferred, so long as the coverage (*q.v.*) is sustained.

Finally we compare the overall accuracy of the estimators, which we can do because in a simulation study the truth is known. The accuracy of the estimator is measured by the root mean squared error (RMSE), which is reported in the units of the data, and smaller is better.

The slippage estimate is formed by subtracting the known number of consignments intercepted by BAU from the estimated number of contaminated consignments, so its statistical properties will be the same as those of the estimate of the contamination proportion (Section E.6.1).

E.6 Results

E.6.1 Pathway contamination estimate

Figure 8 compares the coverage of the three estimators, and presents no red flags. The coverage of the ratio estimator is always higher than nominal, and of the conditional estimator occasionally lower, but not substantially so. Any of the interval estimates would be satisfactory. Figure 9 shows the confidence interval widths, with the conditional estimator being preferred and also showing a lesser impact of decreased sample size in the high-risk stratum.

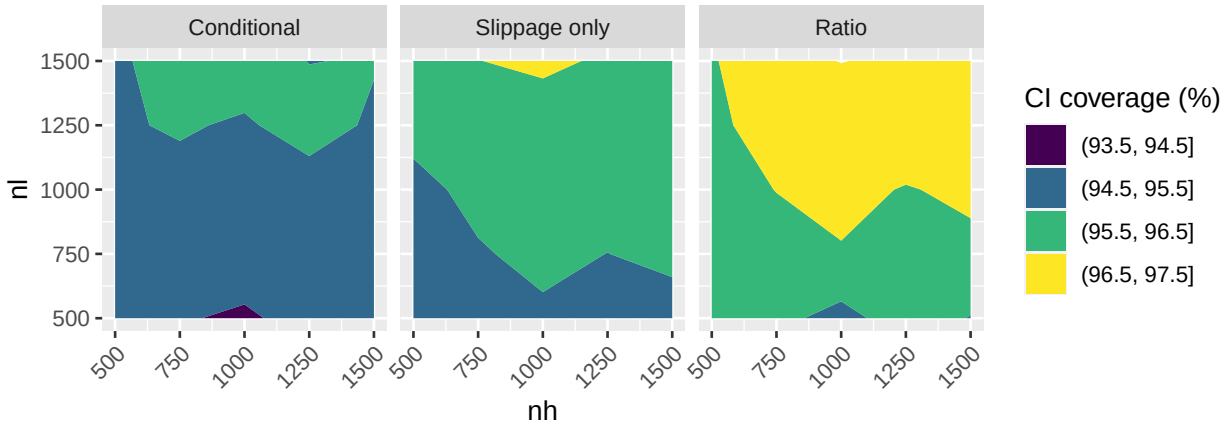

Figure 8: Simulated 95% confidence interval coverage of the contamination estimate for three estimators.

Finally, Figure 10 compares the accuracy of the three estimators as the root mean squared error, scaled to ease comparison, as a function of the sample sizes for the low- and high-risk strata. As could be expected, the RMSE increases from low to high sample size (bottom-left to top-right.) However, the nature of the change varies by estimator, and overall the conditional estimator(left panel) has the lowest RMSE. Curiously, the pattern for the conditional estimator also differs, as the impact of reduction in high-risk samples (right to left) is less compared with the other panels. This is because the high-risk stratum is already inspected, so the informativeness of the high-risk slippage sample is less than that of the low-risk slippage sample.

The overall conclusion of this section is that the conditional estimate of the contamination rate is preferred.

E.6.2 Targeting efficiency estimate

We advocate only one estimate of the inspection efficiency, namely the odds ratio, which is the relative odds that an inspection of a consignment in the high-risk stratum will result in detecting

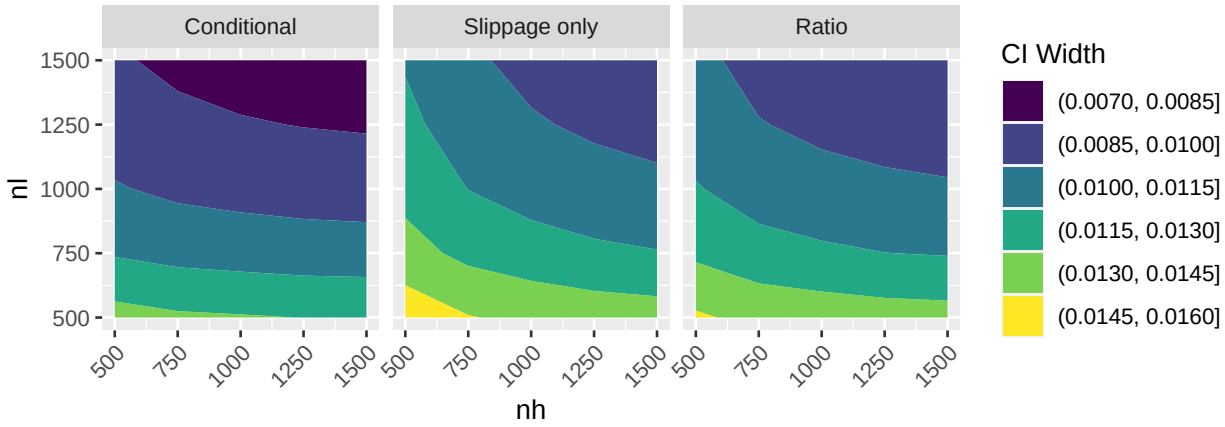

Figure 9: Simulated 95% confidence interval width of the contamination estimate for three estimators, in the units of the contamination rate.

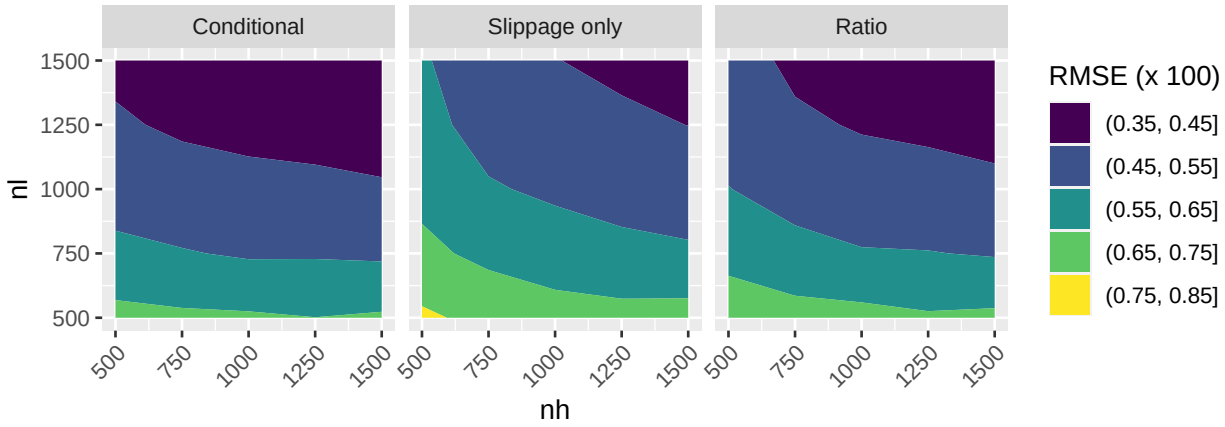

Figure 10: Simulated root mean squared error of the contamination estimate for three estimators.

contamination compared with an inspection of a consignment in the low-risk stratum. It can be thought of as the amplifying effect of the targeting upon the usefulness of inspection.

Figure 11 shows that the confidence interval estimate for the approximation is conservative, with coverage always higher than nominal (which is 95%). This means that the interval estimate is wider than it needs to be. It would be preferable to be exact, but a too-wide estimate is less concerning than a too-narrow estimate. The coverage increases smoothly as the ratio of the sample sizes for the low-risk stratum over the high-risk stratum increases. Figure 12 reports the interval width as a function of the sample sizes, and shows that intervals decrease in width as either sample size increases, which is as expected, with a modest emphasis towards increasing the low-risk stratum sample size as being more influential.

Finally, Figure 13 reports the RMSE of the targeting efficiency estimate shows that the RMSE is mainly a function of the low-risk stratum sample size, across the range of compared sample sizes, with a merely modest effect from the high-risk stratum sample size.

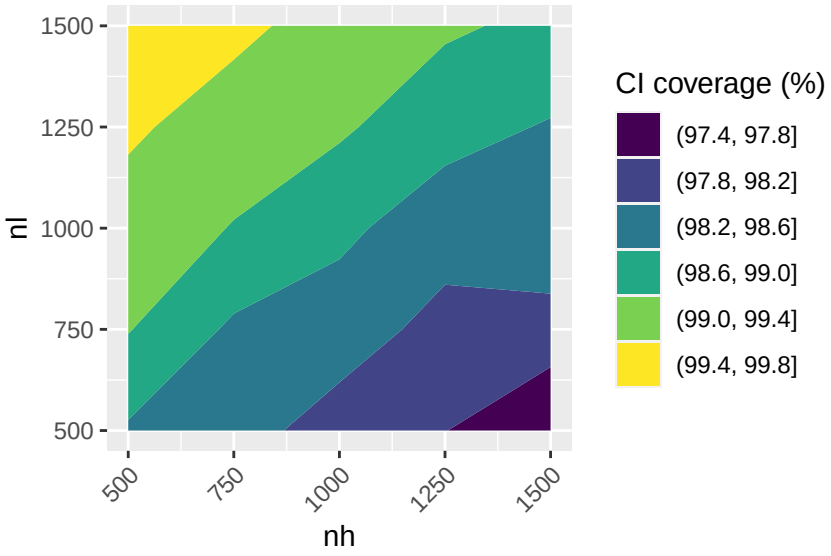

Figure 11: Simulated 95% confidence interval coverage of the targeting efficiency estimate for the approximate estimator.

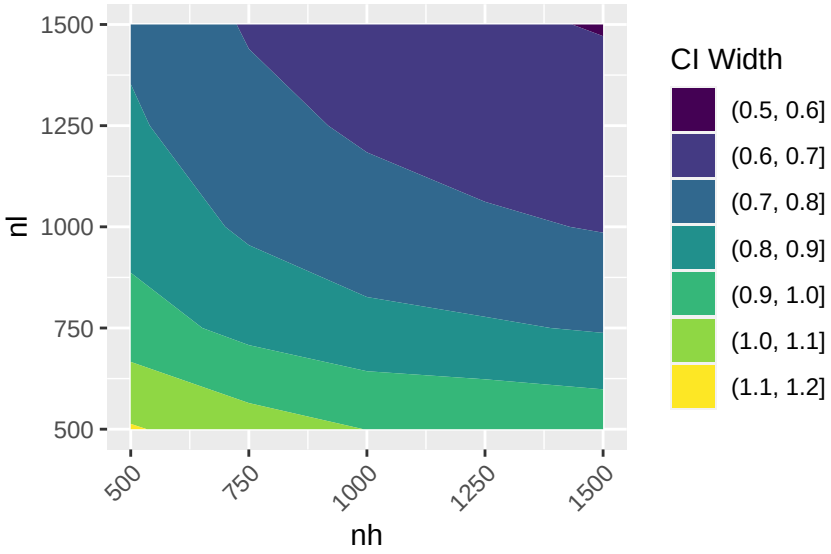

Figure 12: Simulated 95% confidence interval width of the targeting efficiency estimate for the approximate estimator.

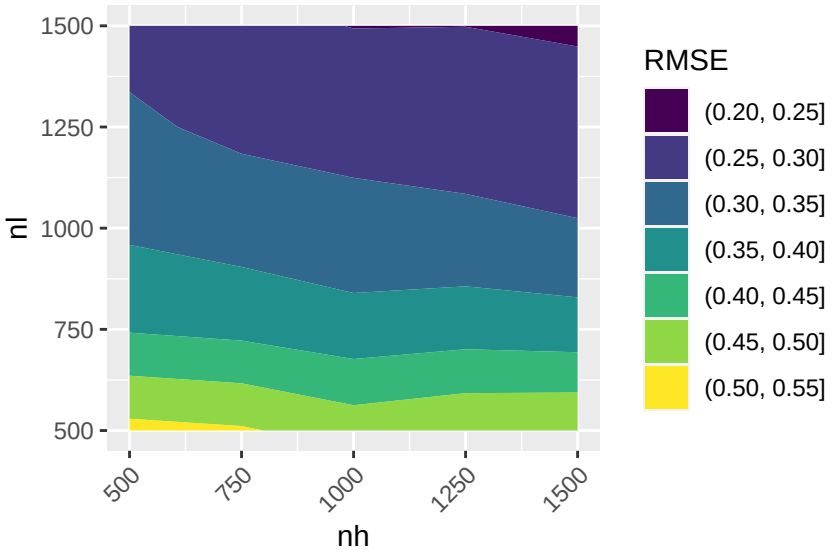

Figure 13: Simulated root mean squared error of the targeting efficiency estimate for the approximate estimator.

The overall conclusion of this section is that the estimate of the targeting efficiency is satisfactory, noting that the interval estimate will be conservative.

E.6.3 BAU inspection efficiency estimate

We advocate only one estimate of the inspection efficiency, namely the probability that the BAU inspection detects contamination that is detected by the slippage survey.

Figure 14 shows that the coverage of the interval estimate of the inspection efficiency is mostly acceptable, but quite low when the high-risk stratum sample size is relatively low. This suggests that the interval estimate has the tendency to be anti-conservative, but probably not to the extent that its interpretation is seriously undermined. Figures 15 and 16 show that the interval estimate width and the point estimate accuracy are entirely a function of the high-risk stratum size, which is as expected.

The overall conclusion of this section is that the estimate of the inspection efficiency is satisfactory, noting that the interval estimate will be anti-conservative.

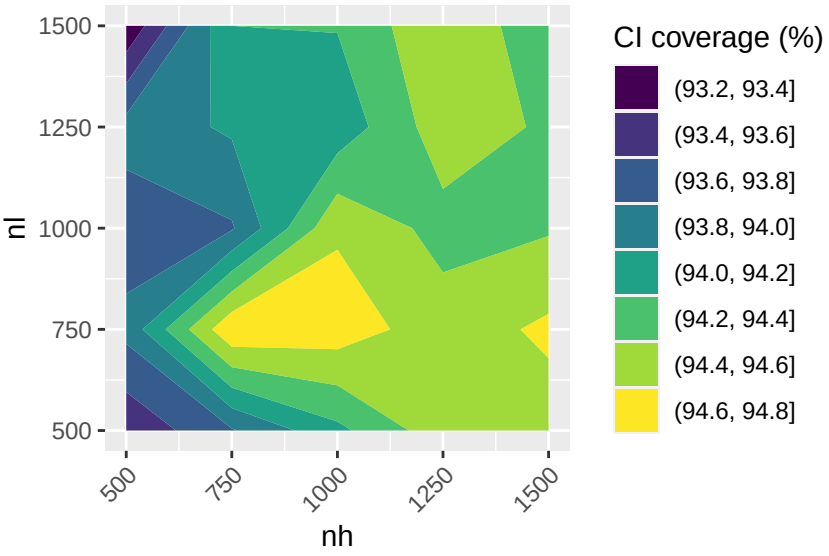

Figure 14: Simulated 95% confidence interval coverage of the inspection efficiency estimate for the approximate estimator.

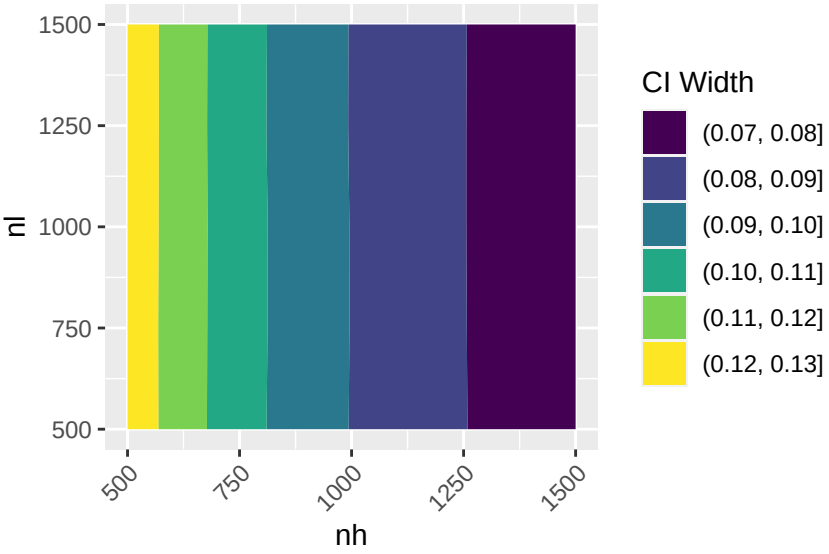

Figure 15: Simulated 95% confidence interval width of the inspection efficiency estimate for the approximate estimator.

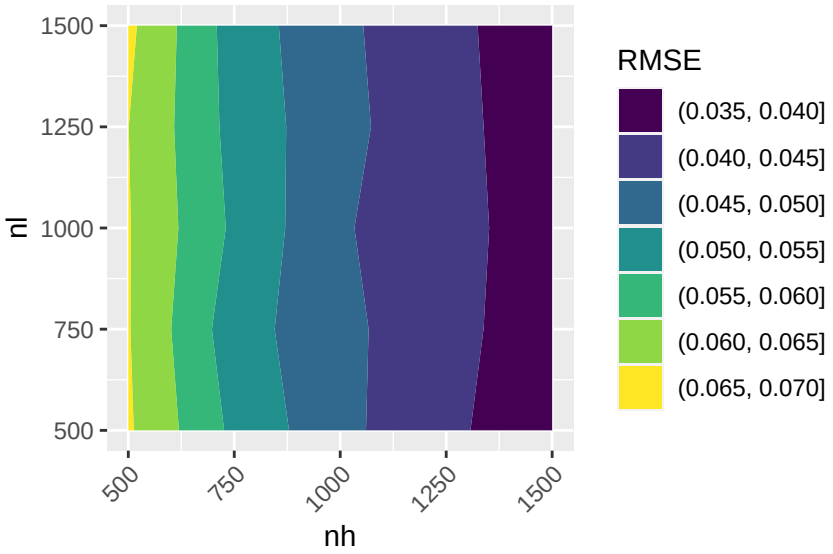

Figure 16: Simulated root mean squared error of the inspection efficiency estimate for the approximate estimator.