CEBRA Project 24D:
Quantitative risk-based prioritisation model for compliance monitoring activities

Version 2.2.0

Andrew Robinson[1], Kara Taylor[2], Anneliese de Vries[2], Nicholas Moran[1], Katja Hohl[2], and Jeremy Walton[2]

[1]CEBRA, University of Melbourne
[2]Department of Agriculture, Fisheries and Forestry

June 18 2025

cebra
Centre of Excellence for
Biosecurity Risk Analysis

Contents

List of Figures

List of Tables

Executive summary

- The Department of Agriculture, Fisheries and Forestry (hereafter, the department) engages in *approved arrangements* (AA's) with operators that allow the latter to manage biosecurity risks in accordance with departmental requirements.
- AA arrangements are established by the *class* of the AA, which imposes certain conditions.
- The compliance of AA's with class conditions is assessed by regular *audits*.
- This project, 24D, reports a collaboration between the department and CEBRA to establish a means by which low-risk (Priority 4) AA's can be prioritised for audits.
- The motivation is to be able to rank P4 AA's in a way that reflects both their likelihood of failure and the potential consequences to Australia's biosecurity.
- This project follows CEBRA Project 23D, which established a quantitative relationship between the frequency of audits and the compliance level of the audited population as a whole (Robinson *et al.*, 2024).
- Phase 1 of the current project assessed the department's data holdings, and found them to be broadly fit for purpose.
- Phase 2 of the project involved fitting statistical models to the department's data holdings.
- The project output is a proof of concept model that reports the modelled risk reduction due to audits (hereafter, audit risk reduction) for the population of P4 AA's at any given time.
  - the risk associated with the AA is taken to be composed of the *likelihood* of failing an audit, which is taken from an update to the model developed in project 23D, and the *consequences*, which are held to be proportional to the number of high-risk entries handled by the AA, that is, the number of entries that require at least one high-risk direction (throughput).
  - the audit risk reduction for each AA is the expected reduction in population-level risk that results from auditing that AA, and also the expected reduction in the risk associated with that AA as a result of the audit.
- The project proof of concept provides interpretable and operationally tractable risk rankings for the population of AA's.

List of Recommendations

1. Introduction

1.1. Background

The aim of this nine-month project was to provide a quantitative method for audit prioritisation for Priority 4 AA's through a risk-based approach to regulatory audits. It follows the CEBRA project 23D "*Quantitative model for assurance-based auditing of approved arrangements*" (Robinson *et al.*, 2024). The current project's objective was to strengthen the risk-based approach and deliver increased confidence that audit resources are allocated and utilised effectively to businesses and activities that pose the highest risk to the department and the Australian community.

Currently, Audit and Assurance Branch (AAB) have purview over 7387 regulated entities that require compliance monitoring through an audit program across Import and Export legislation. AAB developed and adopted an initial risk-based approach (RBA) in 2019 with the intention to better prioritise high-risk audit types within the audit program. The current audit prioritisation is purposely simplistic without a quantitative process. The project required statistical expertise to enable analysis of the department's data holdings and current prioritisation ratings, and to develop a quantitative model for audit prioritisation.

1.2. Approach

The project approach analysed the department's data holdings with the aim of developing a quantitative risk-based prioritisation model that can calculate and assign a priority rating based on risk, considering risk factors across biosecurity, compliance, environmental and intelligence. A two-phase approach for the project was required:

1. Discovery phase
    a) Determine if available data:
        i. were sufficient to build a risk-based model for audit prioritisation, and
        ii. enabled risk trigger identification and integration into a risk-based model.
2. Delivery phase
    a) Develop a quantitative risk-based prioritisation model to provide department confidence that audit resource utilisation is allocated to the highest risk (Chapter 4).
    b) Incorporate risk triggers to enable prioritisation changes where a risk-setting change occurs (Chapter 5).

The project delivered a proof of concept risk-based model for audit prioritisation for Priority 4 AA's. A number of enhancements are possible and were unable to be

delivered in the current project due to time constraints. Here, a proof of concept is taken to mean both in an algorithmic sense, that a workflow of modelling steps can be established that provide a risk ranking for AA's, and in the sense that the models constructed can be fitted with data of acceptable quality, so in a statistical sense.

The balance of the report is laid out as follows. Chapter 2 provides a literature review for methods for risk ranking and prioritisation. Chapter 3 describes the AA and audit system. Finally, Chapter 4 develops a risk-based audit prioritisation model for Priority 4 AA's.

2. Risk ranking & prioritisation methods

Ranking and prioritising a set of options according to their risk has numerous practical applications, including in clinical medicine, supply chain risk, food safety/public health, construction/infrastructure safety, environmental management, climate risk assessment, and so on (Acebes *et al.*, 2024; Austin *et al.*, 2016; Aven & Renn, 2015; Gregory *et al.*, 2012; Heckmann *et al.*, 2015; Ioannou *et al.*, 2022; Van der Fels-Klerx *et al.*, 2018). There are also applications in biosecurity, for example, weed risk assessment, exotic animal disease prioritisation, and border interventions (Brookes *et al.*, 2014; Clarke *et al.*, 2015; Gordon *et al.*, 2008; Bau *et al.*, 2024).

Decision analysis and associated fields like risk assessment and management are relatively new areas of research, with the most common techniques only becoming widely adopted within the last few decades, and new tools and advances continually being developed (Aven & Zio, 2014; Aven, 2016; Gregory *et al.*, 2012). As a result, there is a diversity of methods that may be used, even within specific applications (e.g., see critical reviews of approaches for disease prioritisation, food safety and supply chain risk Brookes *et al.*, 2015; Heckmann *et al.*, 2015; Van der Fels-Klerx *et al.*, 2018).

The appropriateness of any specific method will depend on the requirements of the specific decision context, including the type or amount of data available, the number and diversity of options to be ranked, the time and resources available to the decision maker, and the degree to which stakeholder engagement and values should be incorporated (BIOHAZ, 2012; Dodgson *et al.*, 2009; Gregory *et al.*, 2012; van der Fels-Klerx *et al.*, 2015).

Below, we highlight some general considerations in risk prioritisation problems (Section 2.1), we discuss the specific benefits and limitations of some common approaches (Section 2.2), and describe two specific processes that are commonly used in risk prioritisation (i.e., elicitation and stakeholder preference modelling, Sections 2.3 and 2.4).

2.1. General considerations for risk prioritisation

The risk literature highlights some general issues that should be considered, and specific methodologies often deal with these differently. Some key issues are as follows:

Defining risk & risk metrics: There are various definitions of how risk is both conceptualised and measured (i.e., 'risk metrics'), which may be more or less suitable for different applications. Inconsistencies and ambiguities in how risk and risk metrics are defined within and between studies may undermine the validity and scientific basis of a risk evaluation (Aven & Zio, 2014; Aven, 2016).

Risk is often defined in relation to one or both of: (a) the likelihood of the occurrence of an event, action, or activity; and, (b) the severity of consequences of occurrence (e.g.,

Aven *et al.*, 2018; Ioannou *et al.*, 2022). This is most directly and simplistically applied in risk-matrix approaches (see Duijm, 2015), or simple scoring methods for which risk is estimated as the product of a likelihood and consequence score (e.g., Mansouri *et al.*, 2010, see Section 2.2.1 for further discussion). Multi-criteria approaches often conceptualise risk using a linear utility- or value-function (see Section 2.2.3), where risk is assessed as the weighted sum of a set of risk 'criteria' or 'attributes' (Brookes *et al.*, 2015).

Aven & Zio (2014) highlighted the lack of "well-defined and universally understood terms" in the area of risk assessment and management. The type of 'risk' under consideration is also likely to be specific to the context being considered, so clearly establishing how risk is being characterised and measured is an important initial step in defining the decision context.

Incorporating values & preferences: Many factors beyond technical information may be relevant to characterising risk (e.g., available alternatives to decisions, or subjective perceptions about the values impacted by a potential event or action Gregory *et al.*, 2012; Morgan *et al.*, 2000). Therefore approaches often distinguish between stages of the decision-making process that are fact- versus value-based (Hansson & Aven, 2014). For example, elicitation methods may be distinguished between fact-based technical elicitations and value-based preference elicitations (see Section 2.3), where the latter are used to incorporate the perceptions of decision makers and/or stakeholders into decision-making processes (Bau *et al.*, 2024).

The need for this may vary between contexts. For example, decision trees in clinical medicine often rely primarily on technical data and expert knowledge (Podgorelec *et al.*, 2002; Van der Fels-Klerx *et al.*, 2018), as the outcomes of decision-making processes (i.e., accurate diagnoses) are less reliant on stakeholder perceptions. Whereas in the context of environmental regulation, the *socio-economic theory of regulatory compliance* (Sutinen & Kuperan, 1999) suggests that a number of social factors may be important factors influencing the compliance behaviour of regulated organisations. For example, stakeholders' perceptions of the fairness or efficacy of a regulatory system may influence their level of compliance with that system. This suggests that procedures that incorporate the values of stakeholders may be more effective at managing risk in the context of environmental regulation.

Communication & transparency: Another key consideration is the need to effectively and transparently communicate the steps, reasoning and outcomes of decision-making processes to stakeholders. Different risk-prioritisation methods differ in how easy they are to explain to stakeholders (Van der Fels-Klerx *et al.*, 2018), and the widespread use of simple visual methods like risk matrices in some industries is likely because of their ease of application and interpretation (Acebes *et al.*, 2024). This benefits the decision makers who are applying risk evaluation frameworks but needs to be balanced with other limitations (e.g., the ability of risk evaluation tools to incorporate and assess risk as the result of multiple, competing factors).

Dealing with uncertainty: Several forms of uncertainty are relevant in risk prioritisation, for example, variation in the probability of occurrence or the magnitude of consequences for a future event, variation between expert opinions or stakeholder preferences in elicitations, or incomplete data supporting risk modelling (e.g., Aven, 2016).

Simpler approaches (e.g., risk matrices) have traditionally failed to incorporate and interpret uncertainties in their methods, which may lead to sub-optimal outcomes (Qazi *et al.*, 2021).

Quantitative metrics may characterise likelihood as a probability, including as a probability interval or distribution that incorporates uncertainty (e.g., Aven, 2016). Advanced methods now explicitly incorporate uncertainty, e.g., for expert-elicited data or Bayesian Network modelling (Hester *et al.*, 2024; Weber *et al.*, 2012).

2.2. Benefits & limitations of specific approaches

2.2.1. Risk matrices

Risk matrices are simple 2-dimensional representations of the likelihood and consequences of a potential action or event. Matrices traditionally group risk into qualitative categorical groups (e.g., low, medium, and high). Matrix typologies include entirely qualitative (e.g., for which likelihood and impacts are estimated on Likert scales), semi-quantitative (e.g., as numerical scales, 1–3, etc.), or as quantitative numerical intervals (e.g., low likelihood may be defined as the probability interval: 0.0 – 0.3; Acebes *et al.*, 2024).

As previously mentioned, matrices are relatively common given their ease of application and interpretation, particularly in how information about both axes of likelihood and consequence are clearly represented and communicated to decision makers. Nonetheless, these simpler evaluation methods have been subject to a range of criticisms (see critical reviews in Cox Jr, 2008; Duijm, 2015), including:

- *Uncertainty*, i.e., the use of point estimates and the failure to represent and incorporate uncertainty in risk evaluations.
- *Loss of information*, when aggregating multiple factors into 2-dimensions categorical groupings, and failing to account for situations where competing or interactive factors may influence both likelihood or consequence scores.
- *Biases and subjectivity* in the rating of risks by decision makers when using discrete coloured categories (see for example Proto *et al.*, 2023).
- *Resolution*, i.e., classifying risks into a small number of discrete groups (low, medium, etc.) may not be suitable for contexts requiring prioritisation at finer scales.
- *Design issues*, for example, inconsistencies between the risk levels presented in matrices and the underlying quantitative data that they represent.
- *Applicability*, i.e., matrices may be highly tailored to specific decision contexts, but may have limited applicability across broader and more varied organisational contexts.

Advances in the use of risk matrices have attempted to address some of these issues, for example, continuous consequence diagrams represent the likelihood and consequence scales as continuous axis, and by including interval ratings that incorporate uncertainty on each scale (see Duijm, 2015). Additional advances have also attempted to incorporate uncertainty due to the strength of knowledge supporting ratings or due to variation between experts/risk raters in their ratings (see for example Aven, 2017; Qazi *et al.*, 2021).

2.2.2. Decision trees

Decision trees can be useful tools for grouping large numbers of potential risks into discrete categorical groupings, and require information about only a small number of criteria for each case to determine their risk level (Van der Fels-Klerx *et al.*, 2018). Decision trees are more often used in clinical medicine relative to many other fields, for which statistical analytic techniques can be used with existing data (i.e., a training set) to construct a decision tree for evaluating a set of potential future cases (Podgorelec *et al.*, 2002).

Key benefits highlighted by Batterham & Christensen (2012) and Podgorelec *et al.* (2002) include their hierarchical structure, which readily incorporates interactive or competing risk factors. It is also generally relatively easy to interpret how specific ratings are reached and the factors that were considered in the evaluation. There are also a number of weaknesses in this approach (as per Podgorelec *et al.*, 2002), including:

- *Continuous variables* must be grouped into categorical intervals to be incorporated, and as a result, there may be a loss of information and predictive power.
- *Missing or incomplete data* are difficult to deal with in the design and application of traditional decision trees. Cases with missing data for some categorical factors may not be able to be assessed using the same trees, and these cases may need to be excluded from training sets, potentially leading to biases in the building of trees.
- *Uncertainty* is also difficult to incorporate within the framework, which requires cases to be classified according to discrete criteria.

Nonetheless, these approaches have appeared to perform well at predicting risk when applied in an Australian biosecurity context, with studies using tree-based modelling to analyse air passenger interception data and maritime inspection-compliance data (see Clarke *et al.*, 2015, 2017).

2.2.3. Multi-attribute methods

There is a range of methods for assessing risk as a function of multiple potential risk factors, including systematic methodologies such as Multi-Criteria Decision Analysis (MCDA) and Structured Decision Making (SDM; Dodgson *et al.*, 2009; Gregory *et al.*, 2012). These describe overarching step-by-step processes to make decisions based on an evaluation of some set of possible alternative actions.

Stages of a generalised MCDA process (as per Dodgson *et al.*, 2009) include establishing the decision context (e.g., aims, stakeholders, and methodologies), identifying the options to be evaluated, and determining the objectives and criteria to be used in the evaluation. The scoring of options and weighting of criteria are then estimated and combined to produce an overall value for each option via a range of methods (see further details for the weighting and modelling of criteria under Sections 2.3 and 2.4). This is followed by an examination and sensitivity analysis of outputs.

MCDA has become widely adopted across various fields and commonly uses utility- or value-function approaches to score options based on a linear weighted sum of a set of risk criteria (Brookes *et al.*, 2015; Gregory *et al.*, 2012; Huang *et al.*, 2011), for example:

$$V_i = \sum_{j=1}^{n} c_{ij} w_j$$

Where the 'value' (V_i; or the 'risk level' in the context of risk prioritisation) of each option (i) is a function of a set of criteria values for each option (c_{ij}) multiplied with their criteria weights (w_j; adapted from Brookes *et al.*, 2015). This allows a broad range of inputs to be considered, including both monetary and non-monetary inputs, and allows fine-scale prioritisation of risks (Van der Fels-Klerx *et al.*, 2018).

Alternatives to a linear weighted sum approach also are emerging, such as Bayesian Network (BN) modelling. BN models have the advantage of modelling the distributions and dependencies between criteria, which allows complex/interactive risk factors to be assessed and for uncertainty to be incorporated into the modelling phase (e.g., Straub, 2005; Weber *et al.*, 2012).

A key advantage of MCDA approaches is the ability to assess risks based on a complex set of factors/attributes, using robust and reproducible methods, account for uncertainty, and produce high-resolution rankings of potential risks (Barry & Lin, 2010). As a result, MCDA approaches are highly adaptable across decision contexts.

Another important benefit of MCDA approaches is the explicit incorporation of stakeholder values and preferences in the decision-making process, particularly in determining the importance (i.e., weights) of each criteria (Keeney, 1982). In contrast, monetary-focused frameworks such as cost-benefit analysis or cost-effectiveness analysis are commonly used forms of decision analysis in public sector contexts. These are arguably less useful in environmental regulation and biosecurity contexts as many impacts may not be measured in monetary value, cost data may be missing or incomplete, and these approaches may fail to incorporate stakeholder preferences or values in decision making (Dodgson *et al.*, 2009).

There are also weaknesses of this type of approach, for example, it may be more difficult to communicate how certain decisions were reached for individual cases when compared to simpler risk matrix or decision tree visualisations (Van der Fels-Klerx *et al.*, 2018). These approaches also can be more resource and time intensive, so may not be well suited to decision contexts where a formal analysis is not needed (Gregory *et al.*, 2012; Van der Fels-Klerx *et al.*, 2018). Many approaches also rely heavily on elicitation, and therefore are subject to limitations associated with elicited data (see discussion below). MCDA approaches have also been criticised for taking an overly linear approach to decision making, where the validation and review of weightings/models produced from an MCDA process are not always undertaken (Brookes *et al.*, 2017; Gregory *et al.*, 2012).

2.3. Elicitation for decision making

Elicitation is a component of many risk evaluation processes, for example, to directly rank choices, to estimate criteria weightings, or to estimate the likelihood or potential consequences of a potential event or action. Data may be elicited from experts, decision makers and/or stakeholders to support MCDA processes, and processes can be distinguished between fact-based and value-based (Bau *et al.*, 2024), according to the types of data being elicited, the relevant methods, and their limitations.

Eliciting facts/expert judgements: Where there is inadequate data to support a decision, the judgement of topic specialists may be used as a substitute (e.g., ecologists, entomologists, veterinarians, land managers, industry specialists, etc.; Bau *et al.*, 2024). These approaches focus on estimating 'facts' to support a decision-making process, such as the probability/likelihood of a particular future event or the severity of its consequences. Several protocols exist for expert elicitation (e.g., single round protocols, DELPHI protocols, the IDEA protocol; Hemming *et al.*, 2018), including methods that specifically elicit interval data to incorporate uncertainty into the analysis process.

The key benefit of expert (or technical) elicitations is to provide an empirical basis for decision making in contexts where data is unavailable or incomplete (Ioannou *et al.*, 2022; Van der Fels-Klerx *et al.*, 2018). This is common in biosecurity contexts, where decision makers are unlikely to have pre-existing data on the likelihood or consequences of new pest incursions.

This approach also has weaknesses, as elicited data can be subject to numerous biases. As summarised in Bau *et al.* (2024) (and references therein), this may include cognitive biases (e.g., overconfidence, confirmation bias, anchoring, etc.), motivational biases (e.g., perceived personal gains or losses), and group biases (e.g., highly influential individuals, in-group social pressures, limited diversity in demographic composition or expertise). Elicitations must therefore carefully control for these factors to minimise the effects of the experts' personal values or biases on the validity of elicited data (Sutherland & Burgman, 2015). As a result, it can also be relatively time and resource intensive to prepare an elicitation (Van der Fels-Klerx *et al.*, 2018).

Eliciting values/stakeholder preferences: Value-based approaches are directed towards assessing and incorporating the preferences of decision makers or stakeholders (e.g., industry participants Bau *et al.*, 2024). For MCDA processes, preference elicitation is commonly used for determining the criteria weightings to be applied in weighted sum modelling (Bau *et al.*, 2024; Brookes *et al.*, 2015). A benefit of incorporating preference elicitation into MCDA is that it is inherently participatory, and can be used to engage with and incorporate the perceptions and values of stakeholders directly into the decision-making process (Brookes *et al.*, 2014, 2017).

There are numerous approaches to eliciting preferences (e.g., pairwise comparisons, point scoring, swing weighting; Bau *et al.*, 2024; Brookes *et al.*, 2015; Ioannou *et al.*, 2022). See below for further descriptions and discussion of relevant approaches (Section 2.4). These methods can also share some weaknesses with technical elicitations (especially due to the motivational biases, and sensitivity to group composition). Different approaches also differ in the degree to which they incorporate uncertainty, how they assess trade-offs between criteria, and how easily their processes can be communicated to stakeholders.

2.4. Stakeholder preference modelling

Within an MCDA framework, the weights that are applied to a criteria or attribute set are often determined using elicited preferences from decision makers or stakeholders. Many options are available, which may be suited to different decision contexts (e.g, outranking methods, conjoint analysis, etc. Behzadian *et al.*, 2010, 2012; Brookes *et al.*,

2015). Select approaches include the following:

Direct weighting methods: Directly eliciting criteria weights is common. This is often implemented via point scoring, where stakeholders distribute a limited set number of points (e.g., 100) across a set of criteria, and the average score for each criteria is used (Bau *et al.*, 2024; Brookes *et al.*, 2014). The benefit is that these approaches are easy to implement, and the process for reaching a decision or ranking options can be easily and transparently communicated.

Nonetheless, their limitations include their use of average weights, their failure to consider and incorporate potential trade-offs between criteria that might be relevant to a real decision scenario, and their failure to integrate the scale of criteria used in weighing, which can lead to inconsistencies in the outputs/rankings based on the same weightings (i.e., 'rank reversal'; Brookes *et al.*, 2015),

Swing weighting is a form of direct weighting that addresses some limitations. In this approach, trade-offs between different criteria are specifically incorporated into the elicitation process (for a worked example of this methodology see Bau *et al.*, 2024).

Analytical hierarchy processes: AHP methods are among the most widely adopted forms of preference modelling (Huang *et al.*, 2011). Applications include decision making in manufacturing, engineering, education, supplier/product evaluation, political/governmental decision making, environmental management, and project assessment (Ho & Ma, 2018; Vaidya & Kumar, 2006).

Developed by Saaty (1977), AHP elicits data based on pairwise comparisons between different criteria (Saaty, 2001; Gregory *et al.*, 2012; Bau *et al.*, 2024). Stakeholders rate the relative importance of each criteria on a scale from 1 (i.e., the criteria are of equal importance) to 9 (i.e., criteria A is much more important than criteria B), which are combined to produce weighting values. An extension of this is Analytical Network Processes (ANP) that allows for clustering of criteria (Saaty, 2004).

There is a significant body of literature to guide AHP implementation (Saaty, 2001, 2004, 2016). Another benefit of the wide-adoption of these methods is the development of online platforms and software packages, which support simple implementations of this process (e.g., AHP-OS, ahpweb, an 'ahp' package, etc.; Glur, 2017; Goepel, 2018; Françozo *et al.*, 2023).

Criticisms of AHP/ANP methods have also highlighted several deficiencies. A critical review specifically highlights consistency issues, where limitations to the number of criteria that can reasonably be assessed and traded off by human raters can lead to ranking errors (Asadabadi *et al.*, 2019). Furthermore, Asadabadi *et al.* (2019) suggests that decision makers are particularly sensitive to perceived ranking errors, which can undermine confidence in the decision-making process. Rank reversal between options may also occur if additional options are included in the analysis, even where the scoring of the existing options and criteria weights remain constant (Bau *et al.*, 2024). Although Saaty (2001) argues that may not be inconsistent with decision theory, as the set of available alternatives is relevant to the ranking of existing options.

Probabilistic inversion: Probabilistic inversion is alternative approach to model stakeholder preferences. It has been applied in a more limited range of fields, most notably zoonosis/livestock disease prioritisation (e.g., Havelaar *et al.*, 2010; Brookes *et al.*, 2014)

This approach indirectly obtains criteria weightings by asking stakeholders to rank

a set of constructed risk scenarios, designed to incorporate trade-offs between criteria within hypothetical real-world contexts. Criteria weights are then estimated from the scenario rankings, as opposed to methods that directly elicit the weights or relative importance of criteria. This avoids requiring experts to compare and weigh different attributes that may not have clear physical dimensions or units (Ioannou *et al.*, 2022; Kurowicka *et al.*, 2010). Also, unlike many other approaches, this does not attempt to estimate a single weighting value for each criteria and instead estimates a distribution for each criteria weighting based on the variation in the elicited preferences.

The implementation process for converting scenario rankings to criteria rankings is based on Du *et al.* (2006), and is described in Kurowicka *et al.* (2010) and Neslo *et al.* (2008).

Additional benefits include: allowing for models to be validated based on whether they accurately reproduce the rankings of stakeholders; forcing stakeholders to consider trade-offs concerning more realistic scenarios; and, criteria are assessed as ordinal groupings based on their actual measurement scales, so this avoids assessing criteria based on non-informative ad-hoc scales (BIOHAZ, 2012; Kurowicka *et al.*, 2010; Brookes *et al.*, 2014, 2015).

The main critique of this approach is that its implementation requires relatively sophisticated mathematical techniques, making the processes and rationale behind the outputs more difficult to communicate with stakeholders (Kurowicka *et al.*, 2010). The complexity of implementation also means that simple tools for applying probabilistic inversion to a problem are less likely to be available, relying on technical expertise to implement the process. Like AHP, there are also limitations to the number of criteria that can reasonably be traded off (Brookes *et al.*, 2014). Brookes *et al.* (2015) also suggests that using statistical approaches to estimate weighting may also have larger data requirements, so may not be suited to contexts where only a small group of decision makers are involved.

3. AAs and their audits

3.1. Introduction

This section provides a high-level description of the audit system for approved arrangements (AAs). The policies that guide the auditing of AAs are laid out in DAFF (2023). We provide a brief overview here that is drawn directly from that document.

AAs are voluntary arrangements that are entered into with the department that allow operators to manage biosecurity risks and/or perform the documentary assessment of goods in accordance with departmental requirements, using their own sites, facilities, equipment and people, and without constant supervision by the department and with occasional compliance monitoring or auditing.[1].

Key arrangement outcomes (KAOs) are high level outcomes that the biosecurity industry participant is responsible for meeting under an AA. These are classified activities such as containment, isolation, inspection, treatment, and so on.

The *class(es)* of an AA describe the nature of the arrangement(s), meaning the kinds of actions that the operators are authorised to take. For example, Class 1.1 — Sea and air freight depot (unrestricted) AA sites are approved for "initial non-containerised machinery inspections, rural container inspections, external container inspections and the storage, inspection or treatment of incorrectly certified agricultural products from Khapra beetle countries."[2] An AA may belong to more than one class, and indeed some classes require the AA to belong to other classes.

The compliance of AA's to regulatory requirements is assessed by regular *audits*. These audits involve the assessment of (sometimes, many) *conditions*, the nature of which depend on the classes to which the arrangement belongs. The outcome of the audit (pass or fail) depends on the number and degree of non-compliances against any of the conditions, regardless of the class under which the conditions hold.

Audits are a means by which the department monitors compliance with departmental requirements. Audits are a regulatory assessment of the biosecurity industry participant's ability to meet and maintain relevant AA conditions. Following each audit, the findings are documented, discussed at the exit meeting, and provided as a written report. The report includes details of evidence and associated findings. The result is determined by the severity and number of instances of non-compliance detected by the audit.

The frequency of audits is determined by the risk associated with the operation of the AA. Operators that are newly approved or demonstrate poor compliance are audited more frequently than those which have a demonstrated history of good compliance. AAB established a risk-based approach to prioritise AA sites into one of four categories based on biosecurity activities, commodity type and/or compliance history: Priority 1 (highest), Priority 2, Priority 3, and Priority 4 (lowest).

[1] https://www.agriculture.gov.au/biosecurity-trade/import/arrival/arrangements

[2] https://www.agriculture.gov.au/biosecurity-trade/import/arrival/arrangements/requirements#class-1, accessed 04-Dec-2024.

Risk-based approach (RBA) determinations were introduced in 2019, allowing for AAB to prioritise audit workload to identified AA classes of highest risk through determinations applied with an activity-based and compliance lens ranging from Priority 1 to 3. The Priority 4 framework was introduced in 2024 to be used in conjunction with the existing risk-based RBA determinations to better prioritise risk. This framework categorises audits of AA sites that fall within lower risk classes to Priority 4. These sites will not be audited unless the associated risk level is raised due to one of the following factors:

- a Priority 4 AA site meeting exclusion rules
- a Priority 4 AA site is subject to escalation due to an identified risk trigger(s).

All of these audits are considered in the current statistical analysis.

3.2. Proposed risk amplifiers

As part of the background work leading in to the current project, the department had identified a number of factors that could be considered risk triggers or amplifiers for auditing AAs. The risk factors that had been canvassed by the department are provided in the first two columns of Table 3.1.

The table summarises 18 factors as potential triggers or amplifiers. Codifying all of these factors into a risk model and eliciting weights, which was the original approach we had contemplated, based on the literature review reported in Appendix 2, would have been impossible in the 9 months available for the project, so it was necessary to achieve some simplification.

In thinking about the factors as potential risk amplifiers, it was useful to think about risk as comprising likelihood and consequences, and therefore reflect on whether such amplifiers were likely to enter the risk as *likelihood* or *consequence*. Note that here, likelihood refers to the likelihood that the AA would fail an audit, and the consequences refer to the broader implications of the AA's failure to follow the required requirements. Therefore, for example, we assert that for this exercise, it makes sense for the requirement of high-risk directions or high AA volume to amplify the consequences of an AA audit fail, as opposed to the likelihood of the fail.

Factors that potentially contributed to the likelihood of audit fail that could be detected by an audit were assumed to be covered by CEBRA project 23D, which constructed a model that predicts the probability that an audit would fail if undertaken on an AA, depending on how long it had been since the previous audit and what was the outcome of the previous audit (Robinson *et al.*, 2024).

The role each factor that referred to consequences could play was then either (i) set as an automatic trigger than required an audit, or (ii) reduced to one of a set of 4 more fundamental questions that can capture the risk, namely:

Question 1 Does the AA undertake a sufficiently large number of sufficiently risky directions? (See Section 4.4 for a description of directions).

Question 2 Are the KAOs implied by the AA's class(es) sufficiently risky?

Question 3 Are there any sufficient increases to the volumes of goods handled by the AA?

Table 3.1.: Risk triggers for AAs that might affect audit rates. Triggers are listed in the second column, and the overarching theme in the first. The Role reflects whether the trigger is more likely to affect the likelihood or the consequences. Resolution refers to the questions in the text that capture the trigger; Automatic indicates that an audit is called for directly and 23D refers to CEBRA project 23D (Robinson *et al.*, 2024).

Theme	Trigger	Role	Resolution
Commodity and non-commodity	High-risk goods	Consequences	Q1
	Country of Origin	Consequences	Q1
	Non-commodity concerns	Consequences	Q1
Activities AA undertakes	Treatments	Consequences	Q2
	Inspections	Consequences	Q2
	Verifications	Consequences	Q2
	Waste management	Consequences	Q2
Compliance	Previous compliance history	Likelihood	23D
	Audit history (# audits, last audit date)	Likelihood	23D
	Period since last audit	Likelihood	23D
	Significant changes to AA	Likelihood	Automatic
Environmental	Seasonal risks	Consequences	Q3
	Offshore emerging risks	Consequences	Q3
Intelligence	AA import volumes	Consequences	Q1
	Referrals of non-compliance	Likelihood	Automatic
Political and Reputational	Entities that might be subject to review (external or internal; export space)	Consequences	Q4
	Pathways/markets subject to certification review	Consequences	Q4
	Markets with high inspection or sensitivity (market access sensitivities)	Consequences	Q4

Question 4 Are there any other political or regulatory considerations that contribute?

This reduction substantially simplified the consequences assessment. However, the objective of the current project was to develop a working proof of concept model. In order to develop such a model within the available time, we focused only on the first question, namely the number of risky directions undertaken by the AA, for the consequences. We covered the likelihood contribution of the risk by deploying the modeling approach of CEBRA project 23D.

Development of the likelihood (from 23D) and consequences (Question 1 only) modules and combining them into an overall risk model are presented in the following chapter. Approaches to incorporating the other three questions are provided in Section 4.6.

4. A risk-based audit prioritisation model for Priority 4 AAs

4.1. Objective

Two audit processes are considered, namely (i) audits of AAs brought to the department's attention by various means, and (ii) audits of AAs done to provide assurance that requirements are being met. The primary aim of the project is to identify triggers for the former audits and develop a risk-based schedule for the latter audits. Our goal for this chapter is to develop a risk model for AA auditing that takes account of the inherent riskiness of the AA, for Priority 4 AAs.

4.2. Approach

Cannon (2009) documented a number of ways of allocating audit effort to instances of differing riskiness. For this proof of concept we adopted the "minimise the time until problems are detected" approach, reasoning that for AAs, merely being out of compliance is not necessarily problematic unless they undertake activities that reflect risk, and their inherent risk is therefore arguably proportional to their level of activity, measured in some way.

So, our variation on the Cannon approach is to minimise the number of activities undertaken whilst the AA is in a non-compliant state. Following Robinson *et al.* (2024), we assume that an AA starts in a state which would be declared compliant by an audit, and after a certain amount of time this state may change to one that would be declared non-compliant by an audit, and it remains in that state until audited. The relevant allocation equation is

$$n_i = N \times \frac{\sqrt{\lambda_i \times q_i \times m_i}}{\sum_j \sqrt{\lambda_j \times q_j \times m_j}}, \tag{4.1}$$

(Cannon, 2009, Eqn. 3.6), where n_i is the amount of effort allocated to AA i (e.g., number of audits per year), λ_i is the failure rate (per year, for example, but the unit is irrelevant and simply alters the scale) of AA i, m_i is the throughput (e.g., number of risky directions or entries), q_i the potential consequences, and N is the total resource available.

In the present case, lacking other information we set $q_i = 1$, which implies that the AA's under consideration are held to be equally consequential all else equal. This means that the allocation will be proportional to the square root of the product (multiplication) of the failure rate and the throughput. The time unit represented by λ is irrelevant because we are presenting relative risks (that is, the risk of an A relative to other AAs). We take the relative risk say r_i of the AA at time t since its last audit as being

$$r_{it} = \sqrt{\lambda_{it} \times m_{it}}. \tag{4.2}$$

This value can be used to rank the AAs for auditing, and if the only purpose of the risk measure is to rank the AAs then the square root function is unnecessary. However, the square root function provides a scale that reflects the rate at which resources should be allocated to the audit, hence also providing a relative urgency.

The notion that underpins equation 4.2 is that we are auditing in a steady-state situation for which, at some point in time, the system changes to non-compliant. In other words, we acknowledge that the risks may differ between the various AAs but that they stay the same over the period being considered. Furthermore, the model assumes that we don't have any additional information that the risk may have changed, for example because of Compliance Case Management System (CCMS) referrals. The distribution of our finite amount of resources is chosen to minimise the number of activities under which any problems are undetected. Note that this function is further updated in Section 4.3.1.

4.3. Likelihood: update the Project 23D model

In order to predict the failure rate for an AA, we need to know the probability of audit failure for that AA. Following CEBRA project 23D we can predict the audit failure based on the time since the last audit and whether or not the last audit was a fail (Robinson *et al.*, 2024). The authors constructed a model that predicts the AA population-level audit fail rate from the inter-audit timing.

We could have used the model developed during that project, however the dates of the audits used for model construction ranged from January 2014–September 2023; we now have access to audits completed up to the end of 2024, which is slightly more than 3000 extra in-scope (namely, scheduled or probation) audits (25,772 on 5,122 unique AAs updated from 21,968 on 4,620 unique AAs). We refitted the model proposed in the previous report.

Recommendation 1. *The current project (24D) uses an audit fail prediction model approach that was developed in CEBRA Project 23D. New audit data have become available since the end of 23D, but testing or changing the model with new data was not in the remit of the current project. The department should consider testing or changing the model with the new data (alternative model approaches are canvassed in Section 4.7).*

The relevant statistical model prediction as a function of the relevant predictors is presented in Figure 4.1. The vertical width of the ribbons in the model shows the availability of data; the narrower the ribbon, the more data are available and therefore the predictions of the model are more certain. Here the ribbons show that there is substantially more data for the AA's for which the previous audit was a pass (the green ribbon is much narrower than the red ribbon). This also explains the trumpet-like shape of the ribbons; there are fewer AA's that have two years since their last audit than there are AA's that have 6 months since their last audit.

Figure 4.1 reports the predicted probability that an audit of a random AA (Priority 1-3 or Priority 4) held at time x years after its previous audit would fail given that the previous audit was a fail (red) or not (green), and whether the audit was *announced*

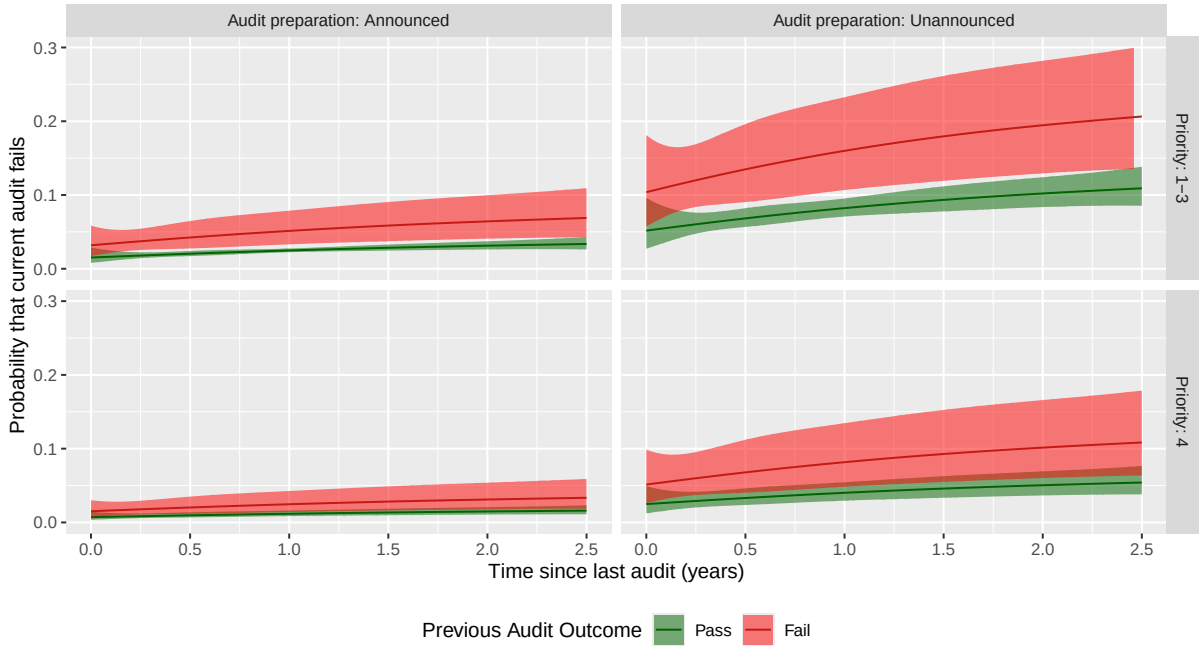

Figure 4.1.: Modelled proportion of audit-level failures, distinguished by outcome of previous audit, audit preparation (announced or unannounced and the AA Priority, updated from CEBRA project 23D. The solid lines show the best prediction and the transparent ribbons show the approximate 95% confidence intervals around the best predictions.

or *unannounced* (called "audit preparation"). The solid lines show the best prediction and the transparent ribbons show the approximate 95% confidence intervals around the best predictions.

Comparing the model reproduced here with that developed in CEBRA project 23D, the failure probability for AA's for which the previous outcome is a PASS are indistinguishable, and the failure probability for AA's for which the previous outcome is a FAIL are about the same as for 23D at the left side (around 0.1) but lower at the right side (0.20 instead of 0.25) however the width of the confidence interval at the right-hand end tells us that data are few and should make us wary to conclude that there has been an important change in the population.

We note that the model output (not shown here) suggests that the odds of audits detecting a fail in Priority 1–3 AAs are substantially higher than detecting a fail in Priority 4 AAs (Odds ratio = 2.1:1, 95% CI 1.6–2.9:1). These results suggest that Priority 4 AAs are substantially more compliant than Priority 1–3 AAs.

We also note that the model output suggests that the odds of detecting a fail by unannounced audits are substantially higher compared with announced audits (Odds ratio = 3.5:1, 95% CI 3.0–4.2:1). Assuming that choosing unannounced audits does not suggest any greater prior suspicion of the audited AA than announced audits, these results suggest that unannounced audits are substantially more sensitive for detecting a fail than announced audits.

Recommendation 2. *Statistical analysis of the audit data shows that the odds of detecting a fail by unannounced audits are substantially higher compared with announced audits. Consequently, unannounced audits may be substantially more sensitive for detecting audit fails than*

announced audits.

Given this model we can then predict the instantaneous failure rate for an audit undertaken at the AA, under assumptions. We can make the greatly simplifying assumption that the occurrence of failures within an AA is a so-called Poisson process (this is the simplest statistical model for these kinds of data). Note that this assumption is probably incorrect and we make it to provide a mechanism for relating failure probability to failure rate.

If that be the case, then we can compute the failure rate for an AA as follows. Let p_{it} be the audit failure probability of AA i at time since last audit event t, which can be taken directly from the model presented in Figure 4.1, either the red line (if the last audit was a fail) or the green line (if the last audit was a pass). The probability of any fail under the Poisson process is as follows

$$\begin{aligned} \Pr(\text{fail}) = p_{it} &= 1 - \Pr(\text{no fail}) \\ &= 1 - e^{-\lambda_{it}}, \end{aligned} \tag{4.3}$$

so

$$\lambda_{it} = -\log(1 - p_{it}) \tag{4.4}$$

4.3.1. But why do we audit?

A notable feature of Figure 4.1 that causes some disquiet is that neither of the two fitted lines (red or green) goes through (0,0). That is, the model asserts that if we were to immediately re-audit even a facility that had passed its previous audit, let alone a facility that had failed and then undertaken rectification, there is still a reasonable probability that the re-audit would detect a fail. This was also observed in the modelling exercise for project 23D.

The intuitively most reasonable explanation is that audits are imperfect detectors of non-compliance, so even if we think that the AA population is entirely compliant, the model claims that we likely have an underlying and undetected failure rate of, say, 4% — about 100 AAs — that could be detected upon re-auditing.

It is reasonable to believe that audits are imperfect and to conclude that spending on audits, which may improve their sensitivity, is an investment that is worth considering. A formal analysis of the re-audit data (that is, the outcomes for the audits that are close enough to one another) may reveal whether it is likely that something changed between the audits.

However this anomaly may be explained, the implications for the current project are important. A challenge with this approach is that we can readily see from Figure 4.1 that auditing the AAs will not decrease the total risk very much. For example, imagine a Priority 4 AA that passed its previous audit a year ago. If we were to audit this AA again unannounced, then at best we could reduce its audit fail rate from say about 0.04 to about 0.03, which fails to even halve the probability of an audit fail.

This observation encourages us to ask the question: why do we audit? That is, what is the change in the population that we hope to create by the auditing exercise? It may well be simply to audit the riskiest AAs, but we consider it more useful to *undertake the set of audits that most reduce the population-level risk*, and that is not necessarily the same thing.

Consequently, we advocate that the risk be set as proportional to the square root of the *reduction* in the expected audit fail rate that would result from performing the audit. That is, the allocation would be as presented in Equation 4.5, replacing λ_{it} by $(\lambda_{it} - \lambda_{i0})$, where λ_{i0} is the expected audit failure rate for AA i directly after its previous audit, to give:

$$r_{it} = \sqrt{(\lambda_{it} - \lambda_{i0}) \times m_i}. \tag{4.5}$$

This replacement $(\lambda_{it} - \lambda_{i0})$ translates into the rates presented in Figure 4.2.

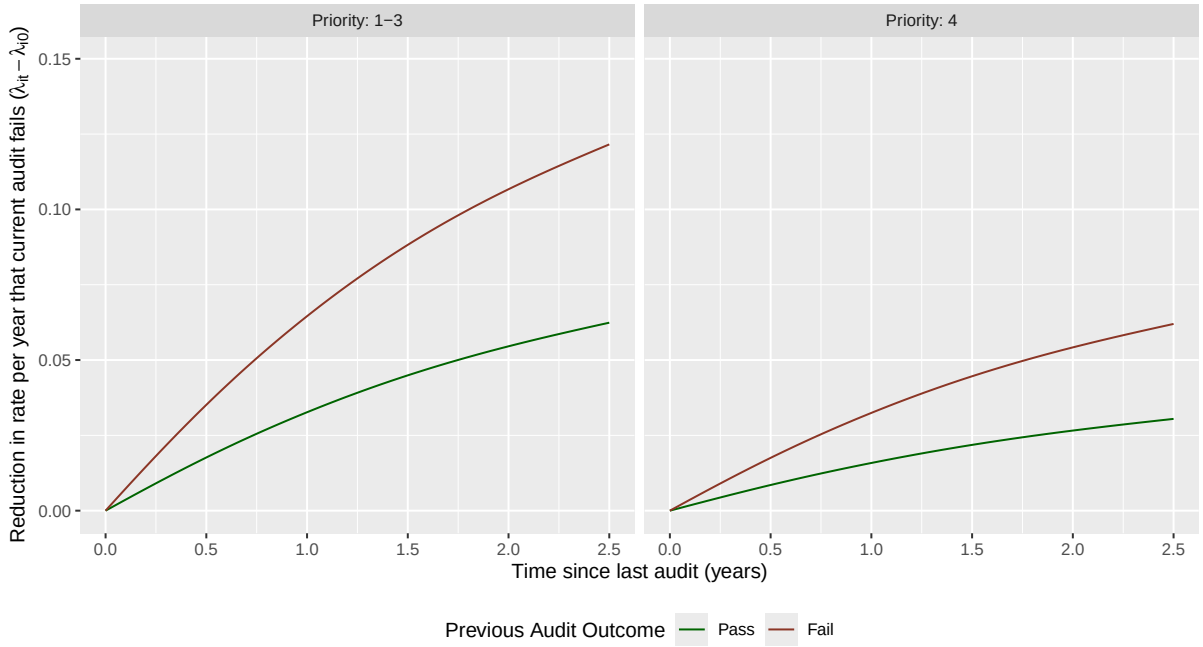

Figure 4.2.: Modelled change in expected audit fail rate arising from unannounced audits, distinguished by outcome of previous audit.

We now apply the model to the AA audit data provided by the department (to be clear, the model was constructed using the full audit histories available for the in-scope AA's, and the predictions are made from the last available audit for each AA). In order to use the model, we need to know the date and outcome of the most recent audit (here we use only previous audits that were *scheduled* or *probation*). Figure 4.3 provides the dates of the most recent audit of any kind for the AA's listed by the department.

4.4. Consequences: count the risk indicators as throughput

A point of clarification is required. Cannon (2009)'s equation, represented here as Equation (4.1), distinguishes between throughput and consequences, allowing the consequences to differ for different units (here, AA's) being non-compliant. We are thinking of the risk as being the product of likelihood and consequences, and do not *a priori* assume that all else equal, different AA's with the same throughput are differently risky (although it would be reasonable to consider the possibility of an AA-class based weight, which is beyond the scope of the current project). In this sense, the consequences for our model map to the product of consequences and throughput in

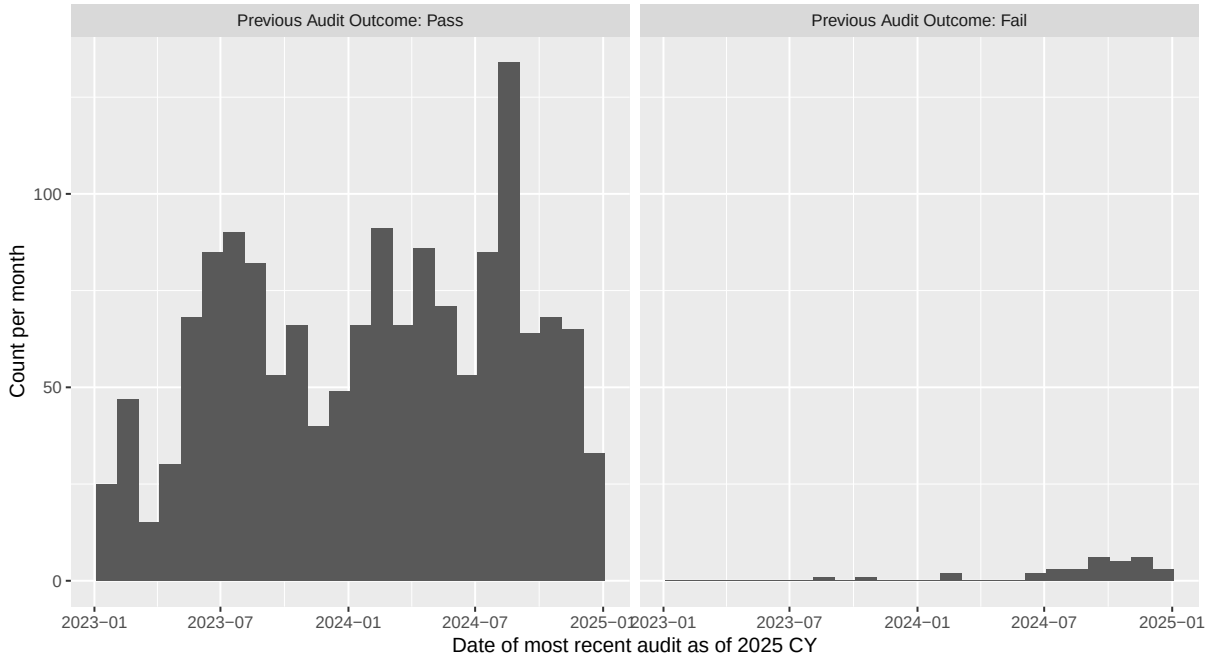

Figure 4.3.: Histogram of dates of most recent audit of any kind for AA's, truncated to January 1, 2023.

Cannon's development. So hereafter we will use the throughput to refer to the consequences.

The department records much of the relevant activity of the population of AAs in the Agriculture Information Management System (AIMS) database. The activities can be captured as *directions*, which are issued to importers by the department as part of mitigating the biosecurity risk for line within the consignment. The department has classified the directions by the level of the risk that they are intended to mitigate.

For this proof of concept we will represent the consequences by counting the number of entries for the AAs that require 'risky' directions. The directions considered risky for this exercise are those among direction categories *Approved arrangements*, *Destruction*, *Devitalisation*, *Disposal Permission*, *Export*, *Fumigation*, *Heat Treatment*, *Irradiation*, *Isolation*, *Movement Allowed*, *Other Treatments*, *Processing*, and *Secure*. Alternative approaches include counting the number of lines that require 'risky' directions and counting the number of 'risky' directions themselves.

So, we will represent the relative consequences of an AA being in a state in which it would fail an audit as being the throughput, where the throughput is the number of entries processed by the AA that entail at least one risky direction. This is in slight contrast to the approach taken by Cannon (2009), who separates the throughput from the consequences.

The entry-counting approach faces a challenge in light of the risk equation 4.5 above, namely that if an AA has no entries that have risky directions then it appears to have no consequences and therefore no risk, which is operationally unattractive. This is a challenge within the data that arises because of the way that the consequences are calculated.

For this exercise we represented the consequences of being in a non-compliant state as the throughput of entries with at least one risky direction for the 2024 calendar year, using AIMS data supplied by the department. A shortcoming of this approach is that

an AA that changed its activity profile during that year may have an artificially low count relative to the other AA's because of a different time of operation. However if the count is low because the AA just started processing risky directions then there will be other reasons to audit (e.g. change of class) and if because the AA has stopped processing risky directions then arguably the previous risk level is no longer met.

A simple work-around would be to set the contribution of the throughput m_i for any AA's that have no risky entries to be equal to (or perhaps slightly below, if possible) the lowest non-zero consequences level for any AA in the same class, which we shall refer to the *class-level baseline*. That is, AA's that would otherwise seem to have no consequences should be set equal to the lowest non-zero consequences among all the AA's in their class. In the case of AA's having multiple classes, we would advocate the highest among the class-level baselines. We did not have time to implement this solution for the current study, so we set the consequences measure for those AA's that had no risky entries to be 1.

Recommendation 3. *This project uses counts of entries with at least one high-risk direction as a means of representing the consequences of audit fail for Priority 4 AA's. This approach fails to capture the risk for a number of in-scope AA's because they have no such entries. The department should consider work-arounds such as allocating a minimum consequence to each AA based on its (riskiest) class, using either (i) the risk of other AA's that share the class as a proxy, or (ii) including risk information about the Key Arrangement Outcomes that are required by the class.*

4.5. Choosing which Priority 4 AA's to audit

The individual components of the AA audit risk reduction are presented in Figure 4.4. The figure shows clustering at predicted rate reduction of 0.1, which we interpret as resulting from the preponderance of previous audits that failed being held within the six months before the end of the time period (right-hand panel, Figure 4.3). There is also a cluster of AA's at 0/1 entries with at least one high-risk direction, which is as expected due to the large number of AA's that had no record of risky directions in 2024.

Figure 4.5 reports the statistical distribution of the relative risk values under the current model. The histogram has a log-scale vertical axis to assist interpretation, and shows strongly that the majority of the AA's are low risk (owing to their low consequences scores as shown in Figure 4.4 but a number of the AAs show appreciable risk-reduction possibilities for auditing.

To provide examples, the highest and lowest 20 AA's by risk reduction are listed in Table 4.1. As may be expected, the low risk-reduction AA's are characterised by low numbers of risky entries (practically, 0) and recent audits that resulted in a PASS. High-risk-reduction AA's have a mixture of high volumes of entries with at least one high-risk direction, times since last audits and outcomes of last audits.

4.6. Discussion

This chapter provides a working proof of concept model that predicts the potential reduction in Priority 4 population risk that would arise from auditing any of the Priority 4 AA's. The proof of concept is mathematical, in the sense that it shows how to link

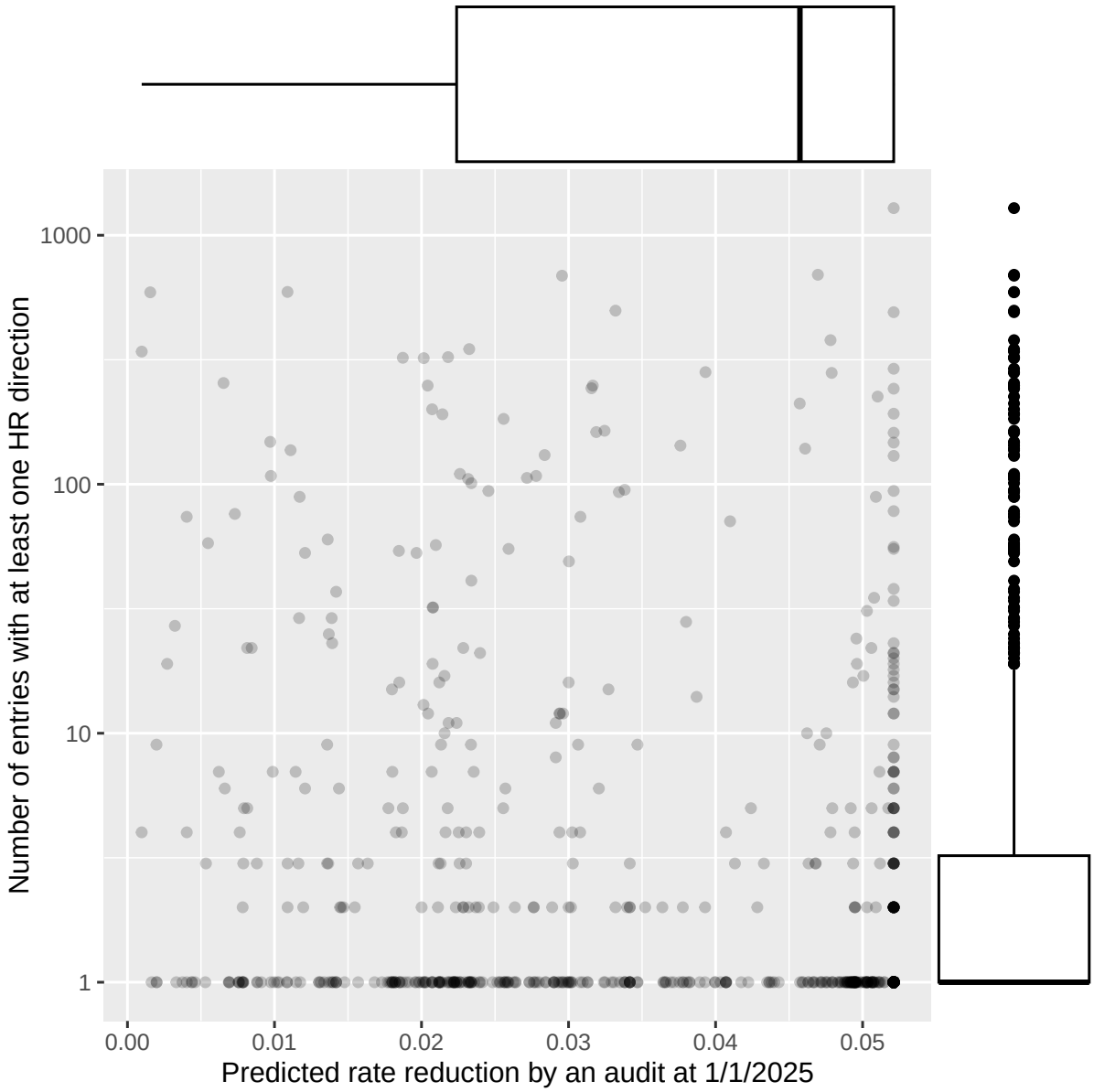

Figure 4.4.: Summary of likelihood and consequences by AA. Marginal boxplots provide the distributions of the horizontal and vertical variables. The vertical axis is plotted on a log-10 scale to ease interpretation.

sources of information to develop a relative risk rating and therefore audit priority for AA's, and a statistical proof of concept, inasmuch as it shows that the models needed to carry out the risk ranking can be sensibly derived from readily available regulatory data, although not without significant technical assumptions. The approach combines departmental information about the likelihood and consequences of audit fails, in a manner that we adapted from approaches laid out in Cannon (2009).

As mentioned in Section 3.2, we focused only on Question 1. Among the other questions, Question 2 could be incorporated by weighted averages using weights established by using Structured Expert Judgement (SEJ); Question 3 by increasing the high-risk direction count to reflect the increase in volume, hence increasing the relative risk of the AA; and Question 4 by the use of guidelines for triggers, perhaps again formulated by SEJ.

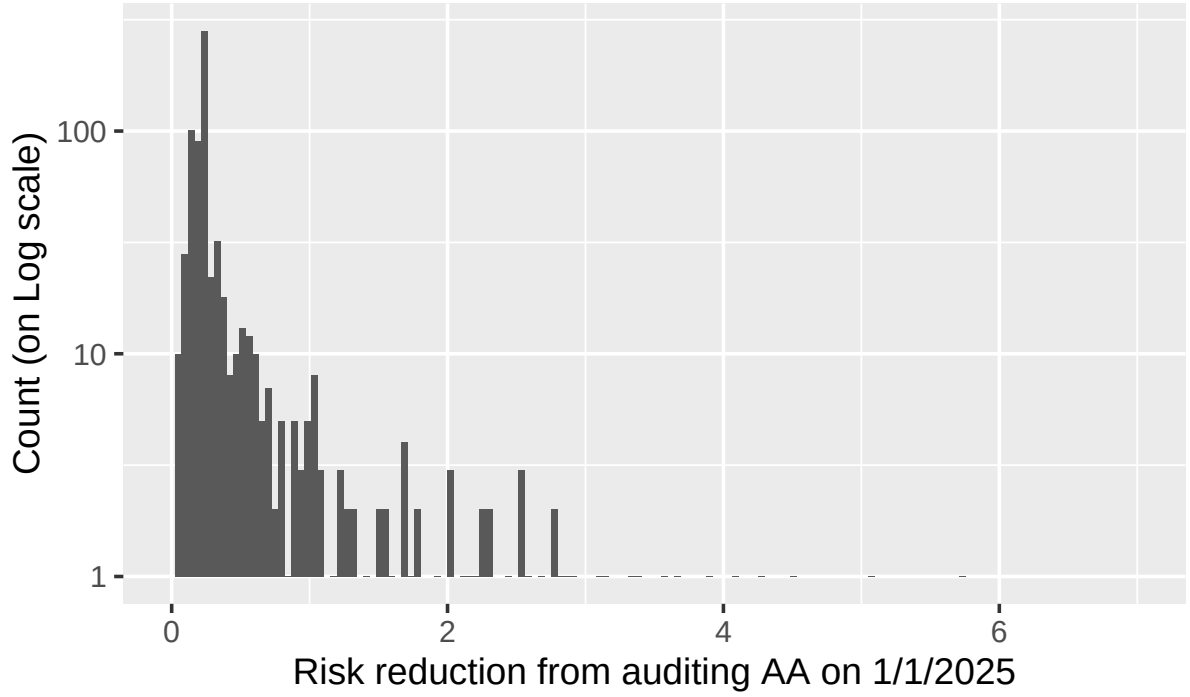

Figure 4.5.: Histogram of audit risk reduction values for Priority 4 AA's, computed at January 1 2025 based on a cap of three years for previous audit age.

We do not consider quantification of uncertainty in the risk estimates in the current model.

4.7. Summary (somewhat technical)

This section provides technical elements that underpin the preceding development and can safely be omitted by the indifferent.

We have a set of AA's (about 3000) that undertake procedures to manage biosecurity risk on behalf of the department. About a quarter of these AA's are considered low priority, for various reasons, and referred to as Priority 4 (P4), which are the only AA's we consider in the rest of this Section. The AA's are audited from time to time, and the audits result in a pass or a fail. The motivating question is: what is the relative urgency with which the AA's should be audited?

Importantly, this is a question that leads to a decision, as opposed to a claim about the population parameters. Our goal is to make a prediction and then use the prediction to make a decision, hopefully a decision that is improved by the prediction, but not necessarily a perfect decision.

We start from the position developed by Robinson *et al.* (2024) that the AA's can be characterised by a binary latent temporal state of being either *compliant* or *non-compliant*. When the AA is compliant then all activities are undertaken in accord with regulatory policies and regulations. When the AA is non-compliant at least some activities do not align with regulatory policies and regulations.

We assume that the compliance state is faithfully detected and corrected by audits. So, immediately after an audit the AA is compliant, and then with a certain unknown rate in time, it will become non-compliant, in which state it remains until the next audit.

Consequently, audits enjoy the function of allowing the regulator to set the expected population compliance level by setting the audit frequency (Robinson *et al.*, 2024).

That previous work distinguished between the urgency of auditing the different AA's based on the time since the AA's last audit and whether the previous audit was a pass or a fail. The AA's were treated as indistinguishable otherwise. Here, we seek to add another variable to set the audit urgency, namely the consequences of the AA being in a non-compliant state. Hence, we start from the position of ranking the AA's by the risk of their non-compliance, where the risk is the product of the rate that they become non-compliant and the consequences of their being non-compliant.

We estimate the non-compliance rate of the AA's by fitting a generalised linear model (glm) to their audit data, where the binary response variable is the audit outcome and the candidate predictors include time since last audit, outcome of last audit, whether the current audit is announced or unannounced, and so on. We obtain a rate from the predicted probability by assuming that the state switch is a Poisson process that happens at a certain rate and is then detected by an audit (if it has happened), hence the binary response variable.

More properly, we note that the response variable is measured at the time of the audit but the state will have changed (if at all) some time before the audit is taken. Therefore the observed duration (i.e., the time between audits) is a left-censored if the audit is a fail (the state change must have happened before the audit, but we don't know when) and right-censored if the audit is a pass (the state change would have happened after the audit, had the audit not reset the process, but we don't know when). Future work could contemplate using survival analysis tools instead of the glm.

Furthermore, in order to apply the glm we need to assume that the fitted relationship from the outcomes of the observed set of audits is representative of fitted relationships of all possible outcomes of sets of audits that might have been taken from the AA's that are in scope. This assumption is worth reflecting upon because unreported factors may have played in to the choice of the AA's to audit. We know, for example, that a number of the audits were carried out because of CCMS triggers, which undermines the representativeness of the sample of audits we have for the overall process. We consider that this is probably not a concern in the current setting because there are comparatively few CCMS-prompted audits. Future work could examine the importance and resilience of this assumption.

Another complication is that the model used to predict the rate at which the AA's become non-compliant does not reset the rate to zero at the time of the audit (Figure 4.1), which leads to the unpleasant possibility that the relative risk of high-consequence AA's may exceed that of sufficiently low-consequence AA's even if the former have just been audited. In order to avoid this possibility we re-scale the rates so that the audit is assumed to reduce the non-compliance rate to zero instead of its lowest possible value (see Figure 4.2). We use this re-scaled rate as the likelihood, but it could be considered more properly the benefit to the non-compliance rate of the population of undertaking an audit of each AA.

Following Cannon (2009), we use throughput to estimate the consequences of non-compliance, specifically, the number of entries processed by the AA that contain one or more high-risk directions, where the risks of all directions were classified by the department. For the current project we assume that the directions are either high-risk or not, but future work may seek to distinguish further between the different kinds of directions.

The risk of the AA can then be transformed by square root to represent the relative urgency of each AA for audit, instead of a mere ranking.

To sum up, we advocate prioritising the AA's by the (square root of) the product of (i) the predicted reduction in AA latent non-compliance rate arising from performing the audit on the AA, and (ii) the number of entries processed by the AA (in, e.g., the last year) that contain at least one high-risk direction.

Table 4.1.: Risk values for the 20 riskiest and the 20 lowest risk Priority 4 AA's based on the current model as evaluated at January 1, 2025. The *Audit date* and *Audit outcome* refer to the most recent audit of any kind (not only scheduled or probation). *Lambda* is the reduction in fail rate, taken from the model represented in Figure 4.2. *HR entries* is the number of entries that include at least one "risky" direction in 2024. *Risk* is the relative rate to which audit resources should be directed to the AA, as computed using equation 4.5. Note that these risk rankings do not involve the improvements suggested in Recommendation 3. AA identifiers have been omitted.

Audit date	Audit outcome	Lambda	HR entries	Risk
11/08/2021	Pass	0.05	1285	8.18
13/09/2022	Pass	0.05	694	5.71
18/10/2021	Pass	0.05	491	5.06
17/08/2022	Pass	0.03	688	4.51
02/08/2022	Pass	0.05	379	4.26
17/02/2022	Pass	0.03	498	4.07
01/10/2021	Pass	0.05	291	3.89
29/07/2022	Pass	0.05	280	3.66
01/12/2020	Pass	0.05	242	3.55
23/02/2022	Pass	0.05	225	3.39
19/07/2023	Pass	0.04	282	3.33
12/08/2021	Pass	0.05	192	3.16
10/11/2022	Pass	0.05	211	3.11
07/09/2021	Pass	0.05	161	2.90
11/09/2024	Pass	0.02	349	2.85
27/02/2024	Pass	0.03	249	2.81
09/05/2022	Pass	0.03	243	2.77
25/02/2021	Pass	0.05	147	2.77
04/07/2023	Pass	0.02	324	2.66
17/05/2021	Pass	0.05	130	2.60
...				
6/07/2024	Pass	0.01	1	0.09
17/07/2024	Pass	0.01	1	0.09
17/07/2024	Pass	0.01	1	0.09
18/07/2024	Pass	0.01	1	0.09
23/07/2024	Pass	0.01	1	0.09
23/07/2024	Pass	0.01	1	0.09
23/07/2024	Pass	0.01	1	0.09
06/08/2024	Pass	0.01	1	0.08
07/08/2024	Pass	0.01	1	0.08
07/08/2024	Pass	0.01	1	0.08
11/09/2024	Pass	0.01	1	0.07
26/09/2024	Pass	0.00	1	0.07
30/09/2024	Pass	0.00	1	0.07
01/10/2024	Pass	0.00	1	0.07
08/10/2024	Pass	0.00	1	0.06
12/12/2024	Pass	0.00	4	0.06
15/10/2024	Pass	0.00	1	0.06
23/10/2024	Pass	0.00	1	0.06
20/11/2024	Pass	0.00	1	0.04
21/11/2024	Pass	0.00	1	0.04
27/11/2024	Pass	0.00	1	0.04

5. Risk factors as triggers

5.1. Introduction

The final exercise for the current project was to establish the benefits of CCMS as a trigger for audits of low-risk AAs (Priority 4). For this exercise we used a dataset that comprised the results for 161 audits undertaken in from October 2023 until September 2024, 18 of which failed (11%). 22 of the completed audits were carried out as a result of CCMS referrals, 6 of which failed (27%).

Statistical details are deferred to Appendix A.1, and are briefly summarised here. The question of the benefit of these triggers can be answered in a statistical framework by using a generalised linear model. The response variable is the outcome of the audit (PASS/FAIL) and the predictor variables are the candidate triggers (e.g., CCMS report) and characteristics of the audits (announced or unannounced).

5.2. Results

Figure 5.1 shows the effects estimates of the final model fitted in the Appendix. Only two terms were retained, namely the CCMS trigger (left panel) and whether the audit was announced or unannounced (right panel). Almost all CCMS audits were unannounced (21/22) whereas about 20% of the verification audits were unannounced (30/139). The 21 unannounced CCMS audits had a higher failure rate (29%) than the 30 unannounced verification audits (10%).

As was determined in CEBRA Project 23D, the audit preparation (namely, announced or unannounced) is material in the current analysis, with the difference ranging from about 3% (announced) to about 10% (unannounced) audit failure rate. The presence of a CCMS trigger was also material with the difference ranging from about 5% (no trigger) to about 10% (trigger) audit failure rate.

5.3. Discussion

The following discussion uses results that are provided in Appendix A.1. It is important to keep in mind that these effects are additive. Therefore, whereas the odds that an unannounced audit will result in a fail, relative to an announced audit, are 3.6:1 (95% CI 0.4–29.9), and the odds that a CCMS-triggered audit will result in a fail relative to a non-CCMS-triggered audit are 3.441 (95% CI 1.1–10.7), the odds that an *unannounced CCMS-triggered* audit will result in a fail, relative to an announced non-CCMS-triggered audit, are 12.6:1 (95% CI 1.3–122.4). The *probability* that an unannounced CCMS-triggered audit will result in a fail is estimated at 0.28 (95% CI 0.13–0.50).

Taken as a whole, these results provide strong evidence that the CCMS triggers are evidence for non-compliance that can be detected by an audit. Note that these results

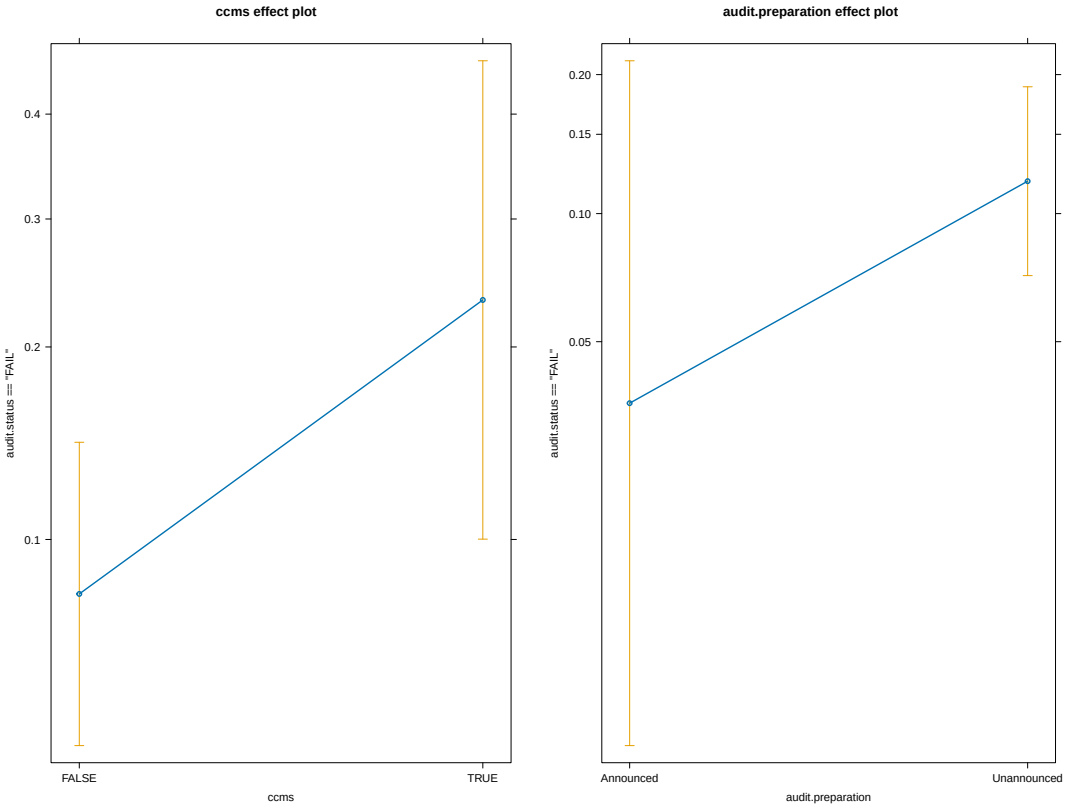

Figure 5.1.: Apparently low-risk arrangement audit outcome model summary. The vertical axis is the estimated probability of an audit fail.

hold regardless of the AA risk as presented in Table 4.1.

Recommendation 4. *Statistical analysis suggests that unannounced audits of low-risk AA's (Priority 4) that are undertaken following non-compliance alerts from the Compliance Case Management System are much more likely to detect non-compliance than audits that do not follow such alerts. Consequently, such alerts should be considered as triggers for automatic audits.*

We also analysed only the verification audits, by which we mean the audits that were carried out without CCMS triggers, in a similar way. Statistical details are provided in Appendix A.2. Broadly speaking, there was no statistical signal in these results, and no guidance concerning the timing of the verification audits can be taken from this analysis. However, it is worth noting that the sample size is small and repeating the analysis in a year or two may lead to more informative results.

Recommendation 5. *Statistical analysis failed to find evidence that the timing of verification audits of low-risk AA's (Priority 4) affects the audit results, which is in contrast to the conclusions of CEBRA project 23D. However, the available sample size for the current analysis was small and these conclusions should be checked after another year of verification audits.*

Bibliography

Acebes F, González-Varona JM, López-Paredes A, Pajares J (2024) Beyond probability-impact matrices in project risk management: A quantitative methodology for risk prioritisation. *Humanities and Social Sciences Communications*, **11**, 1–13. doi:10.1057/s41599-024-03180-5. URL https://www.nature.com/articles/s41599-024-03180-5. Publisher: Palgrave.

Asadabadi MR, Chang E, Saberi M (2019) Are MCDM methods useful? A critical review of Analytic Hierarchy Process (AHP) and Analytic Network Process (ANP). *Cogent Engineering*, **6**, 1623153. doi:10.1080/23311916.2019.1623153. URL https://doi.org/10.1080/23311916.2019.1623153. Publisher: Cogent OA _eprint: https://doi.org/10.1080/23311916.2019.1623153.

Austin PC, Lee DS, D'Agostino RB, Fine JP (2016) Developing points-based risk-scoring systems in the presence of competing risks. *Statistics in Medicine*, **35**, 4056–4072. doi:10.1002/sim.6994. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5084773/.

Aven T (2016) Risk assessment and risk management: Review of recent advances on their foundation. *European Journal of Operational Research*, **253**, 1–13. doi:10.1016/j.ejor.2015.12.023. URL https://www.sciencedirect.com/science/article/pii/S0377221715011479.

Aven T (2017) Improving risk characterisations in practical situations by highlighting knowledge aspects, with applications to risk matrices. *Reliability Engineering & System Safety*, **167**, 42–48. doi:10.1016/j.ress.2017.05.006. URL https://www.sciencedirect.com/science/article/pii/S0951832016306950.

Aven T, Ben-Haim Y, Boje Andersen H, *et al.* (2018) Society for Risk Analysis Glossary. URL https://www.sra.org/wp-content/uploads/2020/04/SRA-Glossary-FINAL.pdf.

Aven T, Renn O (2015) An Evaluation of the Treatment of Risk and Uncertainties in the IPCC Reports on Climate Change. *Risk Analysis*, **35**, 701–712. doi:10.1111/risa.12298. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.12298. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.12298.

Aven T, Zio E (2014) Foundational Issues in Risk Assessment and Risk Management. *Risk Analysis*, **34**, 1164–1172. doi:10.1111/risa.12132. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.12132. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.12132.

Barry S, Lin X (2010) Point of truth calibration: Putting science into scoring systems. *Australian Centre of Excellence for Risk Analysis (ACERA)*.

Batterham PJ, Christensen H (2012) Longitudinal risk profiling for suicidal thoughts and behaviours in a community cohort using decision trees. *Journal of Affective Disorders*, **142**, 306–314. doi:10.1016/j.jad.2012.05.021. URL https://www.sciencedirect.com/science/article/pii/S0165032712003230.

Bau S, Hanea A, Robinson AP, Burgman M (2024) Elicit: Using Structured Elicitation in Biosecurity. In: *Biosecurity: A Systems Perspective*. CRC Press. Num Pages: 20.

Behzadian M, Kazemzadeh RB, Albadvi A, Aghdasi M (2010) PROMETHEE: A comprehensive literature review on methodologies and applications. *European Journal of Operational Research*, **200**, 198–215. doi:10.1016/j.ejor.2009.01.021. URL https://www.sciencedirect.com/science/article/pii/S0377221709000071.

Behzadian M, Khanmohammadi Otaghsara S, Yazdani M, Ignatius J (2012) A state-of-the-art survey of TOPSIS applications. *Expert Systems with Applications*, **39**, 13051–13069. doi:10.1016/j.eswa.2012.05.056. URL https://www.sciencedirect.com/science/article/pii/S0957417412007725.

BIOHAZ (2012) Scientific Opinion on the development of a risk ranking framework on biological hazards. *EFSA Journal (EFSA Panel on Biological Hazards)*, **10**, 2724. doi:10.2903/j.efsa.2012.2724. URL https://onlinelibrary.wiley.com/doi/abs/10.2903/j.efsa.2012.2724. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2903/j.efsa.2012.2724.

Brookes VJ, Barry SC, Hernández-Jover M, Ward MP (2017) Point of truth calibration for disease prioritisation-A case study of prioritisation of exotic diseases for the pig industry in Australia. *Preventive Veterinary Medicine*, **139**, 20–32. doi:10.1016/j.prevetmed.2017.01.017.

Brookes VJ, Hernández-Jover M, Neslo R, Cowled B, Holyoake P, Ward MP (2014) Identifying and measuring stakeholder preferences for disease prioritisation: A case study of the pig industry in Australia. *Preventive Veterinary Medicine*, **113**, 118–131. doi:10.1016/j.prevetmed.2013.10.016. URL https://www.sciencedirect.com/science/article/pii/S0167587713003267.

Brookes VJ, Vilas VJDR, Ward MP (2015) Disease prioritization: what is the state of the art? *Epidemiology & Infection*, **143**, 2911–2922. doi:10.1017/S0950268815000801. URL https://www.cambridge.org/core/journals/epidemiology-and-infection/article/disease-prioritization-what-is-the-state-of-the-art/2BE9E1A0C286ACDC491CA2AF4D3A37E1.

Cannon R (2009) Inspecting and monitoring on a restricted budget—where best to look? *Preventive Veterinary Medicine*, **92**, 163–174.

Clarke S, Hollings T, Liu N, Hood G, Robinson A (2017) Biosecurity risk factors presented by international vessels: a statistical analysis. *Biological Invasions*, **19**, 2837 – 2850. doi:10.1007/s10530-017-1486-1. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85021125193&doi=10.1007%2fs10530-017-1486-1&partnerID=40&md5=ed078c523b31969ff27a5f201d29278c. Publisher: Springer International Publishing Type: Article.

Clarke S, Kuffer A, Hood G, Robinson A (2015) United, we stand: Combining cross-governmental data resources to refine border activities. *Proceedings - International Conference on Data Engineering*, **2015-June**, 87–91. doi:10.1109/ICDEW.2015.7129552. URL <GotoISI>://WOS:000380392100017https://ieeexplore.ieee.org/document/7129552/. Type: Journal Article.

Cox Jr LA (2008) What's Wrong with Risk Matrices? *Risk Analysis*, **28**, 497–512. doi:10.1111/j.1539-6924.2008.01030.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.2008.01030.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.2008.01030.x.

DAFF (2023) Approved arrangements general policies. Tech. rep., Department of Agriculture, Fisheries and Forestry, Canberra. CC BY 4.0.

Dodgson JS, Spackman M, Pearman A, Phillips LD (2009) *Multi-criteria analysis: a manual*. Department for Communities and Local Government: London.

Du C, Kurowicka D, Cooke RM (2006) Techniques for generic probabilistic inversion. *Computational Statistics & Data Analysis*, **50**, 1164–1187. doi:10.1016/j.csda.2005.01.002. URL https://www.sciencedirect.com/science/article/pii/S0167947305000034.

Duijm NJ (2015) Recommendations on the use and design of risk matrices. *Safety Science*, **76**, 21–31. doi:10.1016/j.ssci.2015.02.014. URL https://www.sciencedirect.com/science/article/pii/S0925753515000429.

Françozo RV, Junior LSVU, Carrapateira ES, Pacheco BCS, Oliveira MT, Torsoni GB, Yari J (2023) A web-based software for group decision with analytic hierarchy process. *MethodsX*, **11**, 102277. doi:10.1016/j.mex.2023.102277. URL https://www.sciencedirect.com/science/article/pii/S2215016123002741.

Glur C (2017) ahp: Analytic Hierarchy Process. URL https://CRAN.R-project.org/package=ahp.

Goepel KD (2018) Implementation of an Online Software Tool for the Analytic Hierarchy Process (AHP-OS). *International Journal of the Analytic Hierarchy Process*, **10**. doi:10.13033/ijahp.v10i3.590. URL https://ijahp.org/index.php/IJAHP/article/view/590.

Gordon DR, Onderdonk DA, Fox AM, Stocker RK (2008) Consistent accuracy of the Australian weed risk assessment system across varied geographies. *Diversity and Distributions*, **14**, 234–242. doi:10.1111/j.1472-4642.2007.00460.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1472-4642.2007.00460.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1472-4642.2007.00460.x.

Gregory R, Failing L, Harstone M, Long G, McDaniels T, Ohlson D (2012) *Structured decision making: a practical guide to environmental management choices*. John Wiley & Sons. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781444398557.

Hansson SO, Aven T (2014) Is Risk Analysis Scientific? *Risk Analysis*, **34**, 1173–1183. doi:10.1111/risa.12230. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.12230. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.12230.

Havelaar AH, Rosse Fv, Bucura C, *et al.* (2010) Prioritizing Emerging Zoonoses in The Netherlands. *PLOS ONE*, **5**, e13965. doi:10.1371/journal.pone.0013965. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0013965. Publisher: Public Library of Science.

Heckmann I, Comes T, Nickel S (2015) A critical review on supply chain risk – Definition, measure and modeling. *Omega*, **52**, 119–132. doi:10.1016/j.omega.2014.10.004. URL https://www.sciencedirect.com/science/article/pii/S030504831400125X.

Hemming V, Burgman MA, Hanea AM, McBride MF, Wintle BC (2018) A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution*, **9**, 169–180. doi:10.1111/2041-210X.12857. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12857. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12857.

Hester SM, Bland LM, Arndt E, *et al.*, eds. (2024) *Biosecurity: A Systems Perspective*. CRC Press, Boca Raton. doi:10.1201/9781003253204.

Ho W, Ma X (2018) The state-of-the-art integrations and applications of the analytic hierarchy process. *European Journal of Operational Research*, **267**, 399–414. doi:10.1016/j.ejor.2017.09.007. URL https://www.sciencedirect.com/science/article/pii/S037722171730797X.

Huang IB, Keisler J, Linkov I (2011) Multi-criteria decision analysis in environmental sciences: Ten years of applications and trends. *Science of The Total Environment*, **409**, 3578–3594. doi:10.1016/j.scitotenv.2011.06.022. URL https://www.sciencedirect.com/science/article/pii/S0048969711006462.

Ioannou I, Cadena JE, Aspinall W, Lange D, Honfi D, Rossetto T (2022) Prioritization of hazards for risk and resilience management through elicitation of expert judgement. *Natural Hazards*, **112**, 2773–2795. doi:10.1007/s11069-022-05287-x. URL https://doi.org/10.1007/s11069-022-05287-x.

Keeney RL (1982) Feature Article—Decision Analysis: An Overview. *Operations Research*, **30**, 803–838. doi:10.1287/opre.30.5.803. URL https://pubsonline.informs.org/doi/abs/10.1287/opre.30.5.803. Publisher: INFORMS.

Kurowicka D, Bucura C, Cooke R, Havelaar A (2010) Probabilistic Inversion in Priority Setting of Emerging Zoonoses: Priority Setting of Emerging Zoonoses. *Risk Analysis*, **30**, 715–723. doi:10.1111/j.1539-6924.2010.01378.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.2010.01378.x.

Mansouri M, Nilchiani R, Mostashari A (2010) A policy making framework for resilient port infrastructure systems. *Marine Policy*, **34**, 1125–1134. doi:10.1016/j.marpol.2010.03.012. URL https://www.sciencedirect.com/science/article/pii/S0308597X10000643.

Morgan MG, Florig HK, DeKay ML, Fischbeck P (2000) Categorizing risks for risk ranking. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, **20**, 49–58. doi:10.1111/0272-4332.00005.

Neslo R, Micheli F, Kappel CV, Selkoe KA, Halpern BS, Cooke RM (2008) Modeling Stakeholder Preferences with Probabilistic Inversion. In: *Real-Time and Deliberative Decision Making* (eds. Linkov I, Ferguson E, Magar VS), pp. 265–284. Springer Netherlands, Dordrecht. doi:10.1007/978-1-4020-9026-4_17.

Podgorelec V, Kokol P, Stiglic B, Rozman I (2002) Decision Trees: An Overview and Their Use in Medicine. *Journal of Medical Systems*, **26**, 445–463. doi:10.1023/A:1016409317640. URL https://doi.org/10.1023/A:1016409317640.

Proto R, Recchia G, Dryhurst S, Freeman ALJ (2023) Do colored cells in risk matrices affect decision-making and risk perception? Insights from randomized controlled studies. *Risk Analysis*, **43**, 2114–2128. doi:10.1111/risa.14091. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.14091. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.14091.

Qazi A, Shamayleh A, El-Sayegh S, Formaneck S (2021) Prioritizing risks in sustainable construction projects using a risk matrix-based Monte Carlo Simulation approach. *Sustainable Cities and Society*, **65**, 102576. doi:10.1016/j.scs.2020.102576. URL https://www.sciencedirect.com/science/article/pii/S2210670720307940.

R Core Team (2025) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robinson AP, Moran N, Small N, Whalebone R (2024) Quantitative model for assurance-based auditing of approved arrangements - Final Report. Tech. Rep. 23D, Centre of Excellence for Biosecurity Risk Analysis. URL

https://cebra.unimelb.edu.au/__data/assets/pdf_file/0005/5241587/CEBRA-Project-23D-Report-2.pdf.

Saaty TL (1977) A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, **15**, 234–281. doi:10.1016/0022-2496(77)90033-5. URL https://www.sciencedirect.com/science/article/pii/0022249677900335.

Saaty TL (2001) The Seven Pillars of the Analytic Hierarchy Process. In: *Multiple Criteria Decision Making in the New Millennium* (eds. Köksalan M, Zionts S), pp. 15–37. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-56680-6_2.

Saaty TL (2004) Fundamentals of the analytic network process — Dependence and feedback in decision-making with a single network. *Journal of Systems Science and Systems Engineering*, **13**, 129–157. doi:10.1007/s11518-006-0158-y. URL https://doi.org/10.1007/s11518-006-0158-y.

Saaty TL (2016) The Analytic Hierarchy and Analytic Network Processes for the Measurement of Intangible Criteria and for Decision-Making. In: *Multiple Criteria Decision Analysis: State of the Art Surveys* (eds. Greco S, Ehrgott M, Figueira JR), pp. 363–419. Springer, New York, NY. doi:10.1007/978-1-4939-3094-4_10. URL https://doi.org/10.1007/978-1-4939-3094-4_10.

Straub D (2005) Natural hazards risk assessment using Bayesian networks. In: *9th International Conference on Structural Safety and Reliability (ICOSSAR 05)*.

Sutherland WJ, Burgman M (2015) Policy advice: use experts wisely. *Nature*, **526**, 317–318. Publisher: Nature Publishing Group UK London.

Sutinen JG, Kuperan K (1999) A socio-economic theory of regulatory compliance. *International Journal of Social Economics*, **26**, 174–193. doi:10.1108/03068299910229569. URL https://doi.org/10.1108/03068299910229569. Publisher: MCB UP Ltd.

Vaidya OS, Kumar S (2006) Analytic hierarchy process: An overview of applications. *European Journal of Operational Research*, **169**, 1–29. doi:10.1016/j.ejor.2004.04.028. URL https://www.sciencedirect.com/science/article/pii/S0377221704003054.

van der Fels-Klerx H, van Asselt E, Raley M, *et al.* (2015) Critical review of methodology and application of risk ranking for prioritisation of food and feed related issues, on the basis of the size of anticipated health impact. *EFSA Supporting Publications*, **12**, 710E. doi:10.2903/sp.efsa.2015.EN-710. URL https://onlinelibrary.wiley.com/doi/abs/10.2903/sp.efsa.2015.EN-710. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2015.EN-710.

Van der Fels-Klerx HJ, Van Asselt ED, Raley M, *et al.* (2018) Critical review of methods for risk ranking of food-related hazards, based on risks for human health. *Critical Reviews in Food Science and Nutrition*, **58**, 178–193. doi:10.1080/10408398.2016.1141165. URL https://doi.org/10.1080/10408398.2016.1141165. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10408398.2016.1141165.

Weber P, Medina-Oliva G, Simon C, Iung B (2012) Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Engineering Applications of Artificial Intelligence*, **25**, 671–682. doi:10.1016/j.engappai.2010.06.002. URL https://www.sciencedirect.com/science/article/pii/S095219761000117X.

Revision History

Revision	Date	Author(s)	Description
1.0.0	August 1 2024	NM	Review of multivariate risk-estimation approaches.
2.0.0	April 17 2025	AR	Draft completed with modeling and analysis details.
2.0.1	May 14 2025	KT, AV, KH, JW	Typos identified and fixed; text clarified.
2.1.0	May 27 2025	AR	Corrections to modeling and further corrections to text based on DAFF team feedback.
2.2.0	June 18 2025	AR	Corrections to text based on Scientific Review Panel review reports.

A. Statistical details of the low-risk audit trigger model

The statistical approach used to fit the model that provided the estimates and confidence intervals reported in Chapter 5 was generalised linear modeling (GLM), in which the audits were represented by independent Bernoulli experiments and the logit link function deployed. The output from the near-to-full model that showcases the outcomes reported is presented below. All statistical analyses were carried out using the open-source statistical environment R R Core Team (2025).

A.1. CCMS

The output for the CCMS triggered audit results shows that the statistically significant main covariates in the model were CCMS and audit preparation (using the generous cutoff of $\alpha = 0.05$). Interaction terms were not retained because they were not statistically significant.

```
Call:
glm(formula = audit.status == "FAIL" ~ ccms + audit.preparation,
    family = binomial, data = p4)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -3.472      1.020  -3.404 0.000664 ***
ccmsTRUE                       1.238      0.572   2.165 0.030403 *
audit.preparationUnannounced   1.295      1.060   1.222 0.221573
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 112.79  on 160  degrees of freedom
Residual deviance: 105.42  on 158  degrees of freedom
AIC: 111.42

Number of Fisher Scoring iterations: 6



Analysis of Deviance Table

Model: binomial, link: logit

Response: audit.status == "FAIL"

Terms added sequentially (first to last)
```

```
                  Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                160     112.78
ccms               1   5.2809       159     107.50  0.02156 *
audit.preparation  1   2.0849       158     105.42  0.14876
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A.2. Verification

The output for the verification (non-triggered) audit results shows that there were no statistically significant main covariates in the model (using the generous cutoff of $\alpha = 0.05$). However, the sample size was small, and repeating the analysis with more data in the future may provide useful insights.

```
Call:
glm(formula = audit.status == "FAIL" ~ time_gap + audit.preparation,
    family = binomial, data = verification)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.701921   1.243323  -2.977  0.00291 **
time_gap                     0.003152   0.006549   0.481  0.63032
audit.preparationUnannounced 1.136947   1.072362   1.060  0.28904
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 76.410  on 135  degrees of freedom
Residual deviance: 74.778  on 133  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 80.778

Number of Fisher Scoring iterations: 6



Analysis of Deviance Table

Model: binomial, link: logit

Response: audit.status == "FAIL"

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                135     76.410
time_gap           1  0.15273       134     76.257   0.6959
audit.preparation  1  1.47975       133     74.778   0.2238
```