CEBRA Project 23D
Quantitative model for assurance-based auditing of approved arrangements

Report 2: Phase 2 (Final Report); Version 5.0.0

Andrew Robinson[1], Nicholas Moran[1], Nick Small[2], and Richard Whalebone[2]

[1]Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne
[2]Department of Agriculture, Fisheries and Forestry

December 17 2024

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, department) engages in *approved arrangements* (AA's) with operators that allow the latter to manage biosecurity risks and/or perform the documentary assessment of goods in accordance with departmental requirements.
- The compliance of AA's to regulatory requirements is assessed by regular *audits*.
- This project reports a collaboration between the department and CEBRA to establish a quantitative relationship between audit rate and compliance assurance.
- The motivation is to be able to make a statement of the kind that "an audit rate of x leads to a compliance level of at least y with probability p," where *compliance* is taken to mean that if there were an unannounced audit of all of the AA's, then there is a p probability that y of the AA's would pass.
- The first step of this project was to establish whether the department's data holdings are likely sufficient to the burden of statistical modeling, for example, whether there is sufficient variety in audit rates to permit some useful conclusion.
- Phase 1 of CEBRA Project 23D assessed the department's data holdings for this purpose, and found them to be broadly fit for purpose. The assessment comprised a literature review (Section 2), and a simple summary of the approved arrangement audit outcomes data (Section 3.2).
- The simple summary of the AA audit outcomes data shows considerable variation in audit rate.
- Phase 2 of the project involved fitting statistical models to the department's data holdings and estimating the values needed to produce Figure 1, which links the audit rate to the compliance rate.
- The project output is a model that reports the modelled relationship between audit rate and compliance level, expressed as a probability, enabling statements of the following kind: "there is an estimated 0.9 probability that 94% of the in-scope AA's will be compliant under an audit period of 12 months.", or equivalently, "an audit rate of about 2850 per year leads to a compliance level of 94% with estimated probability 0.9."
- The project results suggest that improved compliance levels may be realised by decreasing the time between audits for AA's for which the previous audit result was a fail, although further exploration of this possibility was beyond the scope of the current project.

Output

Figure 1 reports the modelled relationship between audit rate and compliance level, contextualised by probability. Each line colour represents a different quantile (level) of probability. For example, considering the set of 2850 AA's that were approved at September 1, 2023, there is an estimated 95% probability that at least 2575 AA's would pass an unannounced audit if the audits were every 24 months; this increases by 100 to 2675 AA's if the audits were every 12 months. The right axis reports against the proportion of the population, and the top axis reports the number of audits per year implied by the audit period. Hence we can say, for example, there is an estimated 0.95 probability that at least 94% of the AA's will be compliant under an audit period of 12 months, which corresponds to 2850 audits per year, or reframing the observation in the context of the motivating question, that an audit rate of about 2850 per year leads to a compliance level of at least 94% with estimated probability 0.95.

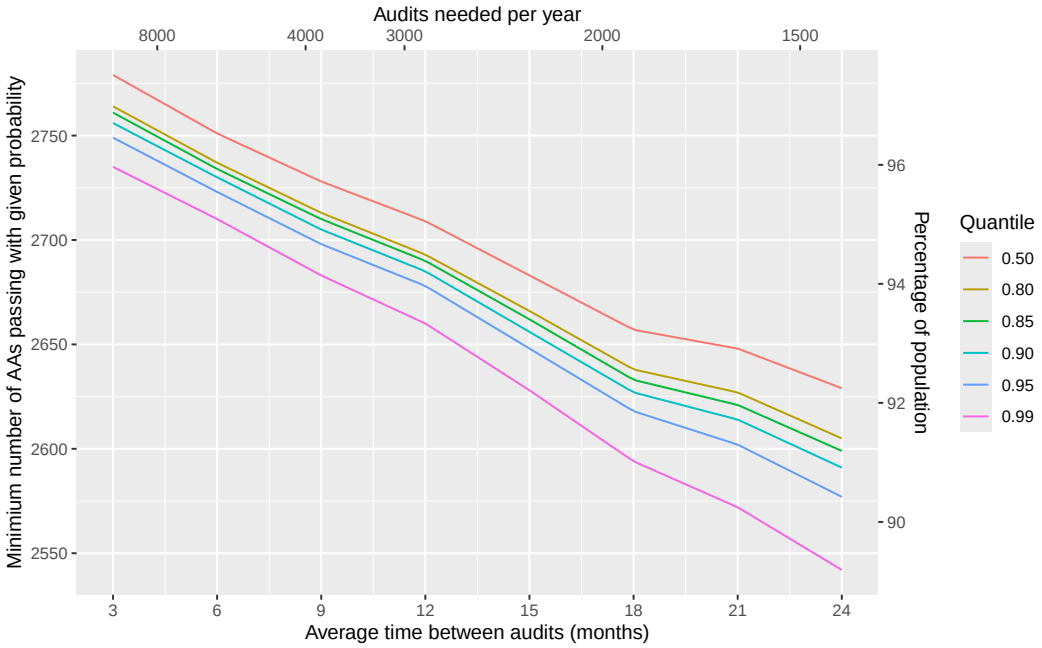

Figure 1.: The lower boundary on the simulated number of AA's that would pass an unannounced audit with given probability as a function of the time between audits, from the 2850 AA's approved at September 1, 2023.

Acknowledgments

This report is a product of the Centre of Excellence for Biosecurity Risk Analysis (CEBRA). In preparing this report, the authors acknowledge the financial and other support provided by the Australian Department of Agriculture, Fisheries and Forestry, the New Zealand Ministry for Primary Industries, and the University of Melbourne. The authors are grateful to Peter Hasson, Pauline Williamson and Erin McElhannan of MPI for very useful conversations early in the project. Detailed review comments from Rob Cannon and Professor David Fox greatly enhanced the final version.

1. Introduction

1.1. Background

It is generally accepted that higher levels of compliance monitoring facilitate higher levels of compliance with performance standards. The broad acceptance of this notion is evidenced in compliance monitoring approaches, policies, and descriptions.

Despite this broad acceptance, however, there appears to have been little if any attention given to exploring whether the relationship can be quantified in a statistically valid manner within the department. Instead, the relationship more often appears to be assumed and the compliance monitoring (e.g., auditing) level set somewhat arbitrarily (e.g., 'one audit per year').

In keeping with this approach, the department's audit frequency regime for approved arrangements, which has been in place for many years, does not have a statistical basis underpinning the audit frequencies that are applied. The department is unaware of any standard, guide or organisation that applies a sound statistical basis to the determination of audit frequencies.

This project will examine whether it is possible to establish a statistical (quantitative) relationship between the frequency of auditing and the level of compliance and, if so, the nature of that relationship. If such a relationship could be established, then it could guide the department in setting auditing levels to achieve chosen compliance levels.

It is not the intent of this project to set audit frequencies but instead to provide a statistical framework which is able to either: (a) set an audit rate to achieve the chosen compliance level; or, (b) determine the level of the compliance rate associated with the chosen audit rate.

1.2. Purpose of this report

Phase 1 of CEBRA Project 23D assessed the department's data holdings for this purpose, and found them to be broadly fit for purpose. The assessment comprised a literature review (Section 2), and a simple summary of the approved arrangement audit outcomes data (Section 3.2). Phase 2 of the project involved fitting statistical models to the department's data holdings and estimating the values needed to answer the motivating question. This report documents the outcome of Phase 2 and conclusion of the project.

1.3. On compliance assurance

The conjectured relationship between audit rate, compliance, and assurance is presented in Figure 1.1. The motivation of the project is to be able to make a statement of the kind that "an audit rate of x leads to a compliance level of at least y with probability

z," where *compliance* is taken to mean that if there were an unannounced audit of all of the AA's, there is a z probability that at least y of the audited arrangements would pass.

As a reviewer (DF) pointed out, it is important to distinguish between the level of compliance and the rate of detection. The former is an attribute of those being audited while the latter is a function of the intensity of auditing. This is a rare case where the two are linked, that is, the intensity of auditing increases the rate of detection (of non-compliance) and also possibly modifies the behaviour, that is, the choice of the auditees to either comply or not comply. This link creates a complex dynamic.

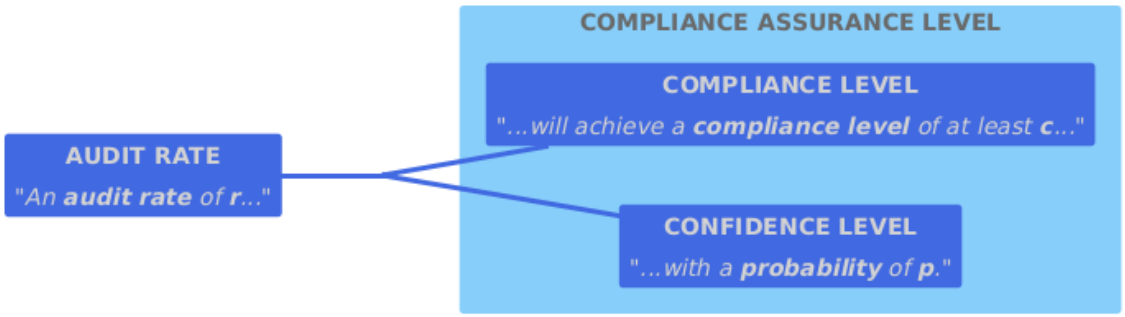

Figure 1.1.: Diagrammatic representation of the potential relationship between audit rate and compliance level, in terms of probability.

1.4. Vocabulary

Approved arrangements (AA's) are voluntary arrangements that allow operators to manage biosecurity risks and/or perform the documentary assessment of goods in accordance with departmental requirements, using their own infrastructure and people, without constant supervision by the department, and with occasional compliance monitoring or auditing. 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)" approved arrangement 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.[1] 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) *criteria*, the nature of which depend on the classes to which the arrangement belongs. The outcome of the audit depends on the number and degree of non-compliances against any of the criteria, regardless of the class under which the criteria hold.

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

1.5. Scope

The project requires the analysis of scheduled and probationary audits for facilities that are approved for classes $1.*$–$8.*$, 9.1, $10.*$–$18.*$, 43.1, and $95.*$, where $*$ is intended as a wildcard, so that $1.*$ matches all classes that start with $1.$

2. Literature review

There has been limited formal research or statistical analysis on the compliance behaviours of industry participants in Australian biosecurity, particularly in response to auditing activities of the department. For example, the CSIRO's 2020 Report "*Australia's Biosecurity Future: Unlocking the next decade of resilience (2020 – 2030)*" recommended investment in social science research "*to better understand non-compliance behaviours*" as a step towards improving industry engagement and shared responsibility in the Australia biosecurity system (CSIRO, 2020).

The purpose of this review is to identify relevant knowledge from existing literature across different fields of research and regulatory compliance, to inform the development of a compliance assurance model.

Specifically in this review we:

- Describe different audit types and relevant terminology (Section 2.1);
- Assess modelling approaches used to analyse audit effects (Section 2.2);
- Summarise the results of studies that have estimated audit effects on compliance across different regulatory areas (Section 2.3); and,
- Identify specific issues relevant to modelling compliance under approved arrangements (AA's; Section 2.4).

2.1. Types and terminology for audits

Audits (or 'inspections') are used in a wide range of fields by regulators to manage the compliance of people or organisations (i.e., 'auditees'). Therefore the scope of this review is intentionally broad, and includes relevant studies from the fields of statistics, economics, public health, environmental regulation, and biosecurity.

Terminology can be specific to regulatory contexts. For example, the term 'audit' is most commonly used in relation to regulatory monitoring of taxpayer compliance (Saw, 2017; Kasper & Alm, 2022). The term 'inspection' or 'environmental inspection' is more commonly used for external monitoring of compliance with environmental regulations (Laplante & Rilstone, 1996; Duflo *et al.*, 2018; Hanna & Oliva, 2010). Although 'audit' may be used interchangeably in the same context (e.g., Telle, 2013), and particularly in the USA, an 'environmental audit' tends to refer to internal compliance-checks within regulated organisations, as opposed to external monitoring actions by regulators (e.g., Evans *et al.*, 2011; Earnhart & Leonard, 2013).

In biosecurity 'audit' is more commonly used in relation to external monitoring of compliance, usually for on-farm biosecurity (Shapiro & Stewart-Brown, 2008; Sandberg *et al.*, 2017; Racicot *et al.*, 2012). Whereas either term may be used in a range of public health and safety contexts, including pharmaceutical quality assurance (Rönninger & Holmes, 2009), workplace OHS compliance (Gray & Mendeloff, 2005; Gray &

Deily, 1996), transport safety (Zarembski *et al.*, 2017; Das *et al.*, 2019; Ivers *et al.*, 2012), medical and nursing home care (Hut-Mossel *et al.*, 2021; Roberts *et al.*, 2022), and food safety regulation (Newbold *et al.*, 2008; Medu *et al.*, 2016).

Audits/inspections across these contexts can all be considered regulatory 'monitoring' actions, which may be distinguished from regulatory 'enforcement' actions such as fines or legal prosecution (Gray & Shimshack, 2011). Therefore for the purposes of this review, we have considered external audits and/or inspections across any regulatory context. Also, we use the term 'audit' interchangeably with 'inspection' to refer to any monitoring action by an external regulatory body to check compliance in a regulated person and/or organisation.

2.2. Audit models

2.2.1. Compliance behaviour modelling

Audits seek to maximise compliance in certain groups, or minimise the losses associated with non-compliance, while subject to soft or hard resource limitations (e.g., time, personnel, or budget constraints). For example, in the context of tax auditing, historically it has not been feasible to conduct annual audits of 100% of taxpayers each year. Therefore, substantial research has been conducted in relation to the strategic planning of audits, in relation to tax as well as numerous other regulatory contexts.

The most directly relevant theoretical approaches for estimating the effects of audits/inspections on compliance levels are based on the Becker (1968) *theory of rational crime*. The core of the theory is that an individual determines their level of compliance as a trade-off between the cost of complying (i.e., C_{comp}), and the potential penalties of non-compliance (i.e., $Pen_{non-comp}$) as well as the probability of their detection (i.e., $Pr(detection)$); per Telle, 2009). This approach assumes that the regulated person/organisation acts rationally to maximise their utility (Sutinen & Kuperan, 1999), such that individuals should decide to comply where the cost of compliance is less than the expected cot of non-compliance, i.e.:

$$C_{comp} < Pen_{non-comp} \times Pr(detection)$$

In this framework, the regulatory tools available to deter non-compliance are adjusting penalties from enforcement actions (e.g., by raising the fines for a exceeding pollution limits in an environmental regulation context), or by adjusting the probability of detecting non-compliance through monitoring actions (e.g., by changing the audit/inspection frequency). Either adjusting monitoring or enforcement levels are the central elements of 'specific deterrence', i.e., where a regulator attempts to deter non-compliant behaviour by a regulated entity through actions directed towards that specific entity (Gray & Shimshack, 2011).

This model has been expanded on and adapted to a number of fields, most significantly in tax compliance (Allingham & Sandmo, 1972; degl'Innocenti *et al.*, 2022), and environmental regulation (Harrington, 1988). Regression-based modelling methods have been commonly used, often to estimate the effects of audit frequency or the actual or perceived probability of an audit occurring (e.g., $Pr(audit)$[1]) on a measure of compliance in regulated entities (Gray & Shimshack, 2011).

[1]Where $Pr(audit) \propto Pr(detection)$ within the Becker (1968) model.

There are considerable empirical studies assessing how enforcement and/or monitoring activities may have positive (and negative) effects on compliance in relation to both tax and environmental regulation (see summaries in Section 2.3 and also Gray & Shimshack, 2011; Traxler, 2014). Therefore this appears to be the most robust and relevant body of research to inform the development of a compliance assurance model.

2.2.2. Alternative approaches

Some distinct approaches to audit scheduling may be found in the areas of *mathematical optimisation* and *game theory* research. In mathematical optimisation studies, the deployment of limited audit resources are often modelled over a planning period with the aim to minimise costs, and may include factors such as the time requirements and costs of different audit activities/auditors, etc. (Dodin *et al.*, 1998; Rossi *et al.*, 2010). While these studies may consider losses that arise in periods between audits, they do not necessarily estimate relationships between audit frequency and compliance itself.

Additionally, *inspection games* are an area of applied game theory, where the relationship between a regulated entity and the regulator is modelled as a non-cooperative game (Avenhaus & Canty, 2009). Originally developed in relation to inspections under the Non-Proliferation Treaty for Nuclear Weapons, this approach has been applied in economic (e.g., tax and insurance auditing), and environmental regulation contexts (e.g., pollution control; see Avenhaus *et al.*, 2002).

A game theory approach has also been applied to biosecurity inspections for imported commodities arriving into Australia, to understand the non-compliance behaviour of importers and design an effective inspection framework (Rossiter & Hester, 2017). Game theory-based approaches may be valuable for understanding compliance levels of auditees under different regulatory systems, and particularly for identifying relevant factors in the auditee-auditor relationship that may be relevant for modelling compliance behaviour. Nonetheless, an approach based on rational behavioural models for compliance (i.e., those based on Becker (1968)) appears to be the most practical and well supported approach for the purposes of this project.

2.3. Audit effects on compliance

2.3.1. Tax compliance

In these subsections we review empirical studies of audit-compliance effects across fields, and identify relevant factors that may also influence these effects.

For tax compliance, several studies have assessed the ongoing effects of audits using real world data. Response variables used in regression analyses may be derived from self-reported income data (e.g., the reported income itself, whether reported income is higher or lower than in the audit year, or the ratio of their reported to actual tax liability; Saw, 2017; DeBacker *et al.*, 2018; Advani *et al.*, 2023). Evidence of greater income reporting following an audit is taken as a proxy for increased compliance. These studies have often found positive effects on compliance measured in the periods following an audit DeBacker *et al.* (2018), although these effects may be more prevalent in groups that were found to have misreported their during the audit (Advani *et al.*, 2023).

This is consistent with an experimental study by Kasper & Alm (2022), which analysed the effects of audit effectiveness and past-reporting behaviour on post-audit tax

compliance. This found that while audits appear to have positive effects on compliance, ineffective audits may have the opposite effect. Experimental studies have also often found that compliance can decrease following an audit (i.e., a 'bomb-crater effect' or 'backlash effect'; Maciejovsky *et al.*, 2007; Mittone *et al.*, 2017; Kasper & Alm, 2022; Advani *et al.*, 2023). For example, post-audit compliance may decrease due to a perceived reduction in the future probability of being audited, motivation to regain lost capital from audits, of perceptions about the efficacy of audits or the audit program. There is also some empirical evidence suggesting that audits may have negative effects on compliance in specific subgroups (e.g., dependent on previous audit outcomes, or charateristics of the auditee; Gemmell & Ratto, 2012).

Therefore, tax compliance research shows some support for positive audit effects on compliance, but also suggests that other moderating factors relating to the audit or auditee may influence the size and the direction of any post-audit effects.

2.3.2. Environmental regulation

There is also a significant body of research into inspection effects on compliance with environmental regulations. The types of compliance responses analysed in these studies may broadly be divided into environmental behaviour and environmental performance measures (Earnhart & Harrington, 2021).

Environmental behaviour can refer to measures of compliance associated with a regulated organisation's actions, often with respect to their meeting of certain regulatory standards or requirements. The actual outcomes of regulatory inspections may be used, such as the number or rate of violations detected in inspections (Eckert, 2004; Telle, 2009). Also, indirect measures may also be used as proxies for compliance. For example, self-auditing is a element of some environmental regulatory systems in the US, and several studies have used the rate of self-auditing as behavioural measure in response to external inspections by government regulators (Evans *et al.*, 2011; Earnhart & Leonard, 2013). Similarly, the adoption of environmental management practices/standards has been used as behavioural proxy (Anton *et al.*, 2004).

Environmental performance instead refers to actual environmental outcomes, for example the rate of pollution by organisations in a certain industry (Gray & Deily, 1996; Laplante & Rilstone, 1996; Telle, 2004; Hanna & Oliva, 2010)[2]. For this project, our focus is on understanding how biosecurity industry participants respond to audits, specifically their levels of compliance with departmental requirements (i.e., 'behavioural' effects), as opposed to determining the actual biosecurity risk that each participant represents (i.e., 'performance' effects).

Generally empirical studies have found positive effects of inspections on compliance, often in relation to pollution/emissions regulation (e.g., Eckert, 2004; Evans *et al.*, 2011; Gray & Shimshack, 2011; Telle, 2013). Although further studies have found no or mixed effects (e.g., Anton *et al.*, 2004; Earnhart & Leonard, 2013). A number of other factors have also been identified that may influence compliance outcomes, including audit quality, the age/size of inspected facilities, social/reputational factors, and the level of cooperation between regulators and regulated parties (Anton *et al.*, 2004; Gray & Shimshack, 2011; Earnhart & Glicksman, 2015; Earnhart & Harrington, 2021).

[2]Although it is important to note that these may overlap for some regulatory systems, for example exceeding a certain threshold of emissions may be considered a compliance failure during an inspection, and therefore a behavioural measure, while also being a measure of environmental performance.

Similar to tax compliance research, there is some positive and some mixed evidence for the effects of inspections on compliance for environmental regulation systems, and evidence that other moderating factors may influence these effects.

2.3.3. Occupational and public health and safety

There appears to be mixed evidence of audit effects for inspections based on health and safety regulations. There is evidence of underlying long-term decreases in OHS compliance across manufacturing facilities in the US, where no-penalty inspections appear to have a particularly negative impact on OHS outcomes (Gray & Mendeloff, 2005). Similarly, Gray & Shimshack (2011) suggests that monitoring and enforcement actions that lack 'teeth', i.e., no-penalty inspections, may lead to negative effects of audits for both OHS and environmental regulation.

For inspections of food premises, several Canadian studies have analysed effects of inspection frequency. In both Newbold *et al.* (2008) and Medu *et al.* (2016), premises were randomly allocated to differing rates of routine food safety inspections. Neither study found evidence that elevated inspections increased compliance, in terms of the number of hazards or infractions detected during inspections. Although it is unclear whether restaurants were aware that they were being subject to increased rates of inspections, and these studies were also conducted over relatively short time frames (i.e., 2 and 3 years, respectively). Also, notably Medu *et al.* (2016) appeared to find increased levels of compliance among all treatment groups over the duration of the study, suggesting inspections may be having some positive effects overall, but the elevated rates of inspection did not provide additional marginal benefits.

2.3.4. Biosecurity

Biosecurity audits also occur in relation to animal production, veterinary or animal-based research facilities, imports and invasive pests and species. Their purpose is often to ensure that biosecurity rules and standards are being followed, to minimise the risk disease spread into and out of facilities (Shapiro & Stewart-Brown, 2008; Porter *et al.*, 2013; Humblet & Saegerman, 2023).

The effects of audits have been tested for on-farm biosecurity. Racicot *et al.* (2012) is a study of chicken production facilities in Canada, which found that experimentally-applied audits failed to improve compliance with biosecurity rules (e.g., hand and shoe washing procedures), although participants were aware that the audits were part of a university study, and no penalties were attached to failed audits. In contrast, Sandberg *et al.* (2017) find an effect of auditing on the prevalence of disease in poultry flocks over a relatively short time period, which is interpreted to be the result of production facilities preparing for audits by ensuring that biosecurity practices were being followed.

Although these studies are a useful starting point for analysing audit effects in biosecurity regulation, there does not appear to be a large body of research specific to audit effects within biosecurity. This highlights the importance of considering approaches from other regulatory contexts to inform the development of a compliance assurance model. Effects will likely vary significantly between regulatory contexts. Nonetheless, reviewing audit effects across contexts is valuable for identifying relevant moderating factors that may influence compliance outcomes (e.g., the strength on enforcement, the effects of the auditees perceptions of their detection risk, audit quality/efficacy, etc.).

2.4. Applying audit models to AA's

2.4.1. Modelling temporal effects of audits (e.g., an audit 'lifespan')

Several approaches have been used across studies to conceptualise and model post-audit effects on compliance over time, which may be relevant to estimating the effects of audit rates on AA compliance. Potentially relevant approaches include the following:

- *The occurrence or frequency of audits* may be used as a predictor in regression models, to estimate the effects of previous enforcement/monitoring actions on subsequent compliance outcomes (see for example, Gray & Deily, 1996; Evans *et al.*, 2011; Earnhart & Glicksman, 2015; Earnhart & Harrington, 2021). The use of lagged measures of compliance relative to when audits occurred is common. The purpose of this is to address the potential perception issues, whereby assessing compliance responses to audits conducted during the same period may not give the audited person/organisation an opportunity to respond to the monitoring effort and increase their compliance behaviour (Gray & Shimshack, 2011). This approach may also partially account for the issue of 'reverse causality' (also described in Gray & Shimshack, 2011). Briefly, monitoring actions are often targeted towards organisations with poor compliance records or that are otherwise considered a high risk of non-compliance. This can lead to statistical modelling that wrongly detects negative effects of increased monitoring effort on compliance levels. Using a lagged measure of compliance relative to the enforcement action may limit this effect, although this may still be an issue when audit actions are targeted and non-compliant behaviour is strongly persistent over time.
- *The predicted probability method* is another approach that may be useful in cases where inspection decisions are targeted towards non-compliant organisations (see for example, Gray & Deily, 1996; Eckert, 2004; Earnhart, 2004; Telle, 2009). These approaches often use two-step regression models, first predicting the probability of an audit (or 'audit threat') based on factors such as the previous audit results/enforcement actions, community or jurisdictional characteristics, or characteristics of the audit themselves. Then the effects of the predicted probability of an audit are estimated (as opposed to, or in addition to actual monitoring/enforcement actions; Gray & Shimshack, 2011). A benefit of this method is that it can account for differences in perceived and actual audit rate effects, which may differ. This may also be well suited to situations where the allocation of inspections is highly heterogeneous and influenced by a range of factors, including past compliance behaviour.
- *The time since the previous audit* may also be used as factor to measure post-audit effects and the time period after an audit (or 'lifespan') within which those effects are observed. These approaches have been used in tax compliance research (see for example DeBacker *et al.*, 2018; Advani *et al.*, 2023). These often use measures of compliance that are separate from the outcomes of audits themselves, allowing compliance effects to be assessed on timeframes independently of when audits actually occur (e.g., to measure the longevity/lifespan of audit effects on compliance). For example, DeBacker *et al.* (2018) found in that US audits increased taxable income and a positive effect persisted for the duration for at least

6 years post-audit. Advani *et al.* (2023) found a similar effect with UK taxpayers, where reported income initially increased in the year after the audit, but this effect steadily declined and disappeared by the 8th year post-audit.

In a theoretical study of tax compliance, Hashimzade *et al.* (2014) considered two forms of post-audit effect, including the 'bomb-crater' effect (described above) and a 'target' effect. This referred to the case where "*subjective probability is increased after an audit and decays when no auditing occurs*", based on the concept that proximity to a recent audit may lead to a temporarily elevated perception of $Pr(audit)$, leading to higher compliance. The concept of audit effects having a lifespan has also been observed in other contexts. For example, Hut-Mossel *et al.* (2021) reviewed how audits improve the quality of hospital care, and established that "*externally initiated audits created quality improvement awareness although their impact on improvement diminishes over time*".

In the context of AA's, the expected frequency and perceived probability of an audit within probationary and regular regimes has been relatively constant over time (e.g., "*one regular audit will be conducted on any business day within any 365-day period*" under low rates for regular audits; DAFF, 2023). So for regular audits, the rate and the AA's perceived probability of receiving an audit may be relatively constant over time.

This suggest that the occurrence, frequency or perceived probability of audits may not be strong predictors for estimating compliance effects, unless there are consistent long-term differences in the audit-rates of AA's that is independent of audit regimes. Furthermore, comparing compliance under higher (probationary) or lower (regular) audit rates is not a viable method for testing audit frequency effects on compliance, due to the different consequences of non-compliance under the different regimes (i.e., $Pen_{non-comp}$ is not constant).

An approach using the time since the previous audit as a predictor may be useful, provided there is some variation in the inter-audit time period in the AA data. Nonetheless, any approach will still need to account for potential effects of targeting, or differences between probationary and regular audits.

2.4.2. Estimating population-level compliance

Studies often model individual compliance behaviour as a binary choice, i.e., an organisation is either compliant or not during a particular time period (e.g., Gray & Deily, 1996; Eckert, 2004; Evans *et al.*, 2011). Regression models are then able to estimate their probability of compliance, and how this varies due to certain characteristics (e.g., an organisation's size, type, age, etc.), which may be used to directly infer the proportion of compliance/non-compliance across a broader population of regulated entities.

Models for compliance based on individual choice have nonetheless been subject to some criticism. For example, theoretical models of individual taxpayer behaviour have tended to underestimate population-level compliance and estimates of compliance based on real world data (Hashimzade *et al.*, 2013). This suggests that factors beyond individual utility may affect compliance decisions. Therefore, conceptual models have expanded to include broader moral, social and contextual factors that may also influence an individual's decision to comply. For example, the *socio-economic theory of regulatory compliance* (Sutinen & Kuperan, 1999) considers a class of 'moral obligation and social influence' factors, such as an individual's reputation, or their perception

of the system's fairness or efficacy. These factors may explain observed patterns that depart from individual models. For example, increased rates of audits could lead to reduced compliance if they undermine confidence in the regulatory system, or groups may exceed regulatory requirements due to reputational benefits of compliance.

Social network effects are often linked to the transfer of information, where an individual being audited may influence an other's subsequent compliance decisions (e.g., Kasper & Rablen (2023) distinguishes 'own' versus 'cross' post-audit effects). For example, Hashimzade *et al.* (2014) assessed taxpayer compliance in an agent-based simulation model, where 'social custom' effects may influence compliance decisions and information can transfer within a social network (i.e., their perceived probability of being audited, if they had been audited, and if they were compliant). As a result, perceived $Pr(audit)$ tended to be higher than the actual rate, leading to higher compliance. Compliance levels also differed between subgroups, due to interactions between social effects and individual risk preferences. Bloomquist (2011) also considered 'neighbourhood' effects in tax compliance, via both an agent-based model and laboratory experiments. They also found that social effects appear to increase reporting compliance, and that individuals generally overestimated their likelihood of being audited.

Network modelling approaches may not be directly applicable to this project, as the available audit/compliance data for AA's may be used to create a statistical model that directly estimates compliance levels and the effects of audits. Furthermore, information transfer between AA's may not be an major factor influencing their compliance decisions, given the level of transparency and information transfer between the department and auditees (e.g., of expected audit rates, consequences of non-compliance, etc.). Nonetheless, these studies do show how compliance levels may differ between subgroups of regulated organisations (e.g., where social or reputational factors differ between industries). This emphasises the need to account for potential grouping factors in models (e.g., the type or identity of the AA).

2.5. Conclusions

Modelling approaches to analyse audit-compliance effects are relatively common, particularly in tax and environmental regulatory compliance. The available literature for biosecurity audits is limited relative to those more developed regulatory areas. Nonetheless, analysing effects of biosecurity audits on compliance is viable and will benefit from incorporating concepts and knowledge from multiple other fields.

Regarding the central question of this project, i.e., can a statistical model be developed to estimate the effects of audit rates on compliance levels for AA's, a behavioural modelling approach appears to be the most practical and well supported approach. Regression modelling is common in regulatory compliance research, and models may include various factors associated with organisation's identity, their regulatory-compliance history, and their social environment. Predictor variables may be based on audit frequencies, or time periods between audits, depending on the level of temporal variation in the dataset. The power of this approach is limited by the degree of temporal variation in the rate of inspections for AA's, which is independent of other factors such as the audit regime (e.g., probationary vs regular), or risk-based targeting. Nonetheless, this modelling approach is consistent with existing literature and may be viable for estimating compliance levels across AA's as a function of audit rates.

3. Approved arrangements audit data holdings

3.1. Description

The department provided CEBRA with extensive AA audit data holdings (Table 3.1, Figure 3.1). The data are taken from the Quarantine Premises Register, which is an IT system used to record approved arrangement types, biosecurity industry participant's details, audit results and what biosecurity directions have been activated for a site. The table AIMS.refDirection was also provided by the department but not used

An initial analysis of these data was undertaken to verify whether there would be sufficient variation in the between-audit timings to be able to assess whether there is any impact on the between-audit timings and the audit outcome. That is, if two audits of a single AA a_j and a_k are time t apart, then can we detect any correlation between the length of t and the outcome of audit a_k? Summary graphics are presented in the next section.

Table 3.1.: Summary of the databases that comprise the dataset provided by the department. Other than the first row, all tables were sourced from the Quarantine Premises Register (QPR); this prefix is omitted here. The Primary Key refers to the relational database identifier of the table.

Table	Rows	Columns	Primary Key
aimsRefDirection	490	16	
arrangement	14630	33	Arrangement SK
arrangementStatus	73652	12	Arrangement status SK
audit	95243	17	Audit SK
auditCriteriaResult	5055160	7	
classStatus	90305	14	Class status SK
correctiveActionRequest	11226	22	CAR ID
directionClassCombination	9357	13	
directionStatus	394996	12	Direction status SK
refAuditCriteriaResult	30936	9	Criteria result SK
refClass	139	8	Class code
refClassConditions	6790	11	Class conditions SK
refDirection	488	16	

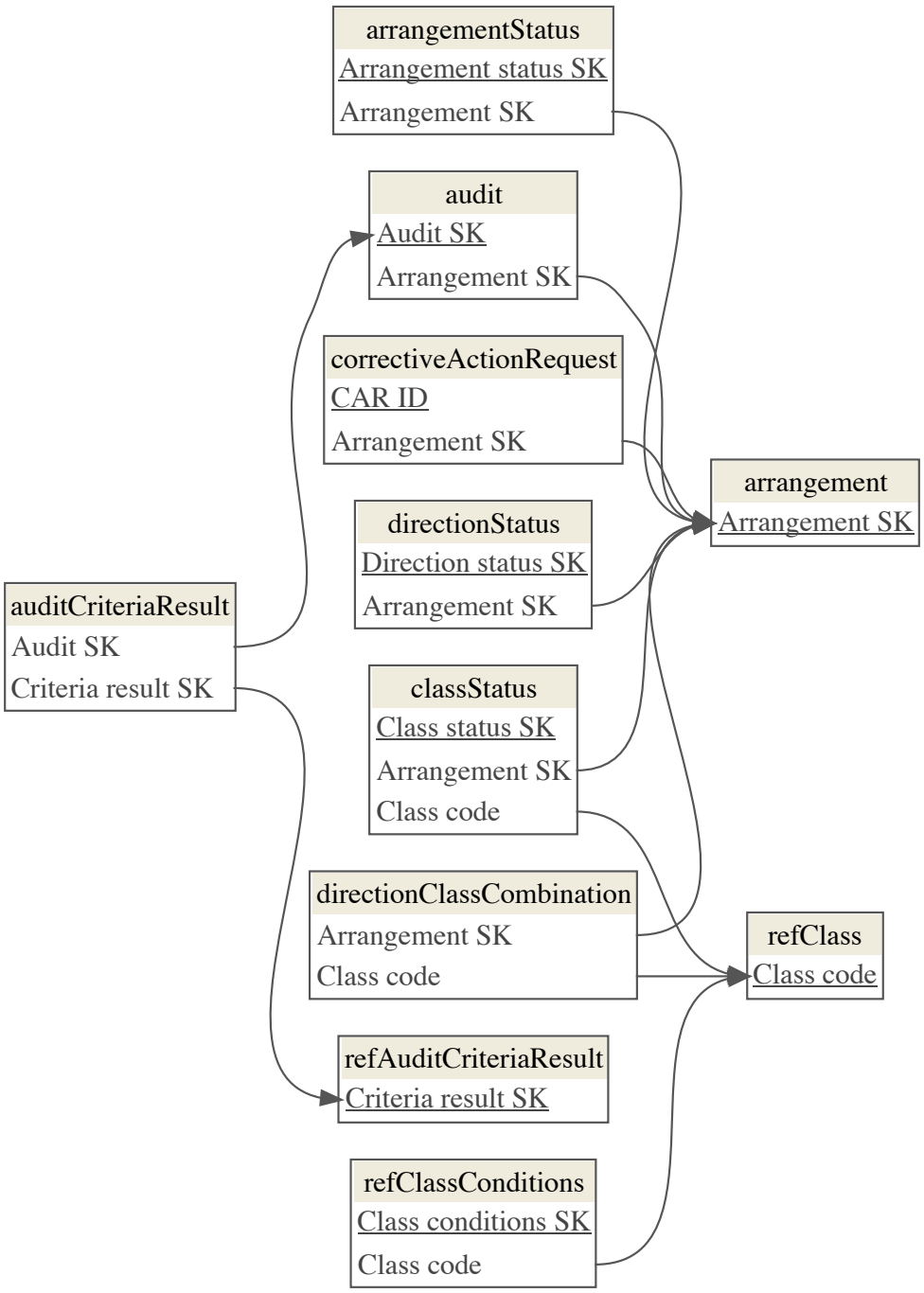

Figure 3.1.: Diagrammatic representation of the structure of the relational database constructed from the department's data holdings. Only the keys for the database are presented here; primary keys are underlined. Each table shown was sourced from QPR.

3.2. Data analysis

This section provides graphical summaries that offer insights into the data in general and the variation of the inter-audit time-spans. For the provided audit data, the scheduled date range of audits was from May 27, 1999 to October 4, 2023. The completion dates of the main three kinds of audits are presented in Figure 3.2.

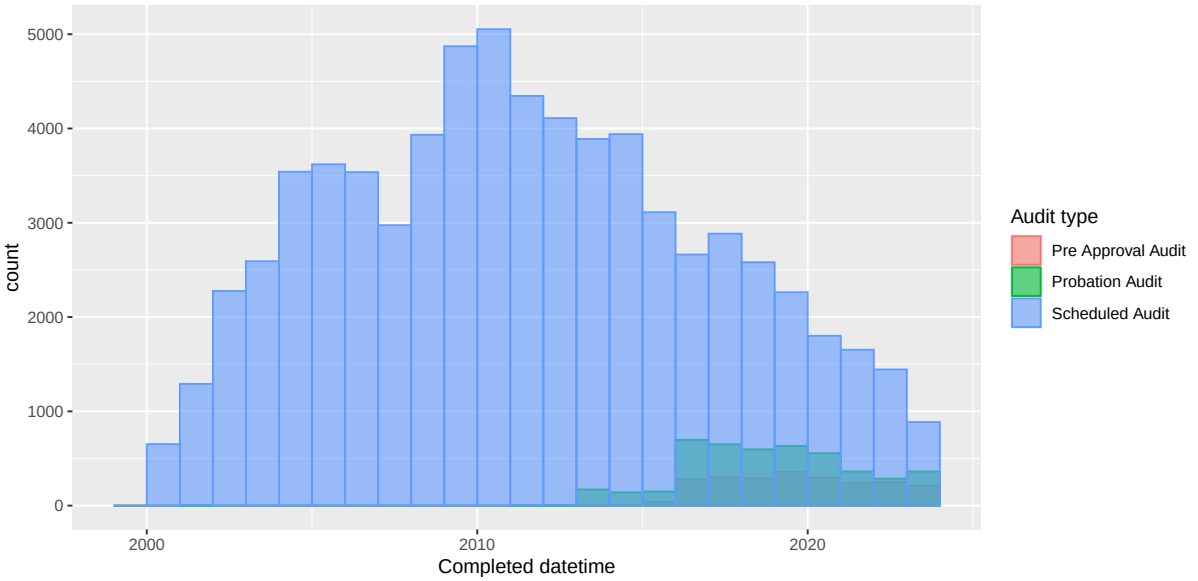

Figure 3.2.: Histogram of completed audit dates by year of completion and audit type. The largest total is Scheduled audit (70,624), followed by Probation (4,620; starting 2014) and Pre-approval (2,276; starting in 2016).

Figure 3.3 provides a compact summary of the key information concerning the scope of the data to answer the motivating question. The data selected for plotting are the second and all subsequent audits for an AA. The x-axis (horizontal) reports the time since the previous audit. The rows are distinguished by whether the audit was announced or unannounced (this is called Audit preparation in the data). We see that there is considerable variation in this value, ranging from several weeks to several years, an outcome that provides (informal) confidence that there will be enough (informal) leverage for a statistical model.

The following list provides pointers that are relevant to future work that involves analysis of the AA audit data holdings.

1. Some 'Initial assessment' records are not the first record. Before May 2013, 'Initial assessment' was used to both: (i) initialise a new AA in the electronic system, and (ii) record the results of pre-approval audits. In May 2013, a 'Pre Approval Audit' type was added to QPR (see Figure 3.2) and is subsequently used to record the outcome of pre-approval audits (where required). From May 2013 onwards, the proper usage of 'Initial assessment' to initialise the AA in the electronic system at the time the application for a new AA is reviewed. Wherein, they are not a record of an audit. While this is the proper usage, it should is noted that some entries under 'Initial assessment' post-May 2013 may be the results of pre-approval audits that have incorrectly been recorded as initial assessments.

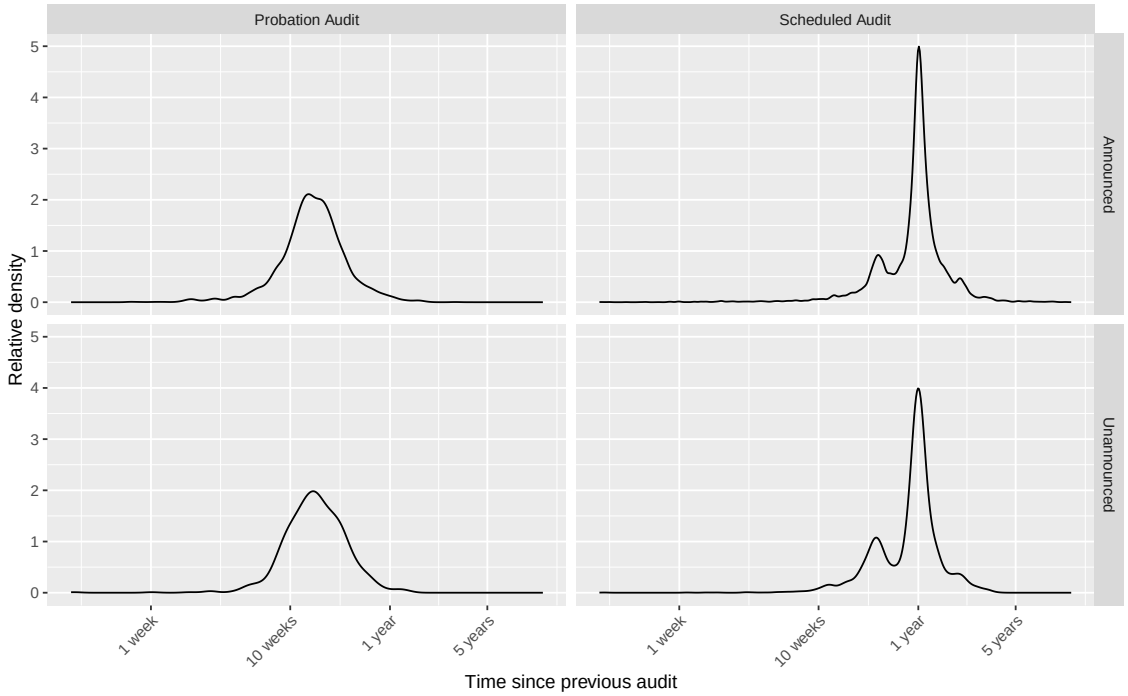

Figure 3.3.: Density of the relative frequency of times since last audit, categorised by audit preparation (announced or unannounced) and type. NB: Pre-approval audits are not included here as all of these audits must have preceding audits.

2. Some 'Pre Approval Audit' records are not the first record — these are for class changes. It is also noted that there are a number of AA's with more than one 'Initial assessment'. This also appears to represent incorrect usage.
3. 'Pre-approval' audits may or may not be 'full-scope' audits. A pre-approval audit for a new AA will always be a full-scope audit, i.e., will assess compliance with all requirements of the proposed AA. A pre-approval audit where, say, an additional class of activity is being added to an existing AA may (a) just assess compliance with the additional requirements of that class or (b) be a full-scope audit.
4. Need to handle 'Non-audit events', which is often just a chance detection of non-compliance e.g., by a biosecurity officer conducting a goods inspection at the AA. These may have useful non-compliance insights (although presumably negatives are not counted, which reduces the value of this analysis within the exercise).
5. Audit outcomes needed work. Specifically, concerning QPR,[1]
    a) From 2000 to around mid-2013, there are no audit results in the [Audit].[Audit status] field. Instead, it seems that auditors (sometimes) put the audit result in the free text [Audit].[Audit comment] field. Accordingly, up to around mid-2013, the [Audit].[Audit status] field contains NULL values.
    b) From 2000 to around March in 2014, non-compliance outcomes against specific criteria are indicated by a value of 'Criteria Unsatisfied' in the [Criteria result] field. These non-compliance outcomes are valid even though there is a value of 0 in the corresponding [Non-conformity] field. Auditors

[1]Unless otherwise noted in [*table*].[*field*] format, all references in this comment are to the [RefAuditCriteriaResult] table.

sometimes indicated whether a non-compliance was minor/major/critical in the free text [AuditCriteriaResult].[Audit criteria comment] field. At other times, there was a comment in this field but not one indicating whether the non-compliance was minor/major/critical. Often, however, this field was left empty, generating a NULL value.

c) From around March 2014, auditors seem to have stopped using the value of 'Criteria Unsatisfied' in the [Criteria result] field. Instead, where there was a non-compliance, they started using (a) 'Non Conformity, Minor' (b) 'Non Conformity, Major' and (c) 'Non Conformity, Critical' as the three non-compliance value options in the [Criteria result] field. Note also that, where the [Criteria result] field contains one of these three values, the [Non-conformity] field always contains a value of 1 (c.f. the earlier situation where the [Criteria result] field contained a value of 'Criteria Unsatisfied').

3.3. Phase 1 conclusions

Feasibility Based on the simple analyses of the department's data holdings presented in Section 3.2, we believe that it is feasible that a statistical model could be constructed to assess the impact of inter-audit duration on audit outcome. Hence we believe that it is possible to estimate and asses a quantitative relationship between audit rate and compliance level. Further work needs to be done by CEBRA on these records to ensure that they are fit for purpose, but the data will likely be sufficient.

Data requirements At this point we do not believe that more data will be required for Phase 2.

Complications Other factors may influence the level of compliance (and whose effects could potentially obscure the influence of audit rate). Such factors could include AA class, audit type, auditor. We can account for these other factors in a statistical model.

4. Statistical modeling

This chapter documents the development, fitting, and interpretation of a statistical model that is used to resolve the question that motivates the project. The technical details are provided in Appendix A.

4.1. Introduction

The goal of this modeling exercise was to develop a model that would enable the prediction of the probability of achieving a given population-level compliance rate as a function of the time between audits and possibly other explanatory factors. The core of such a model is a statistical regression model that can be used to predict the outcome of an audit of an AA as a function of various explanatory factors, for example, the classes held by the AA being audited, the time since the last audit, the outcome of the last audit, and so on.

4.2. Conceptual model

Here we briefly describe the conceptual model that guides our subsequent modelling choices.

1. We assume a population of N facilities that provide a range of services.
2. The facilities can be in one of two latent states, namely *compliant* and *non-compliant*, and the state can only be detected by an audit.
3. We will assume that audits are unannounced.
4. The facilities are subjected to an audit cycle of period r years to determine, and if necessary rectify, their compliance status. Therefore the between-audit duration is $1/r$. Each year, $n \approx N/r$ facilities are audited.
5. We assume that audits will always detect and rectify the non-compliant state, so that facilities are always compliant directly after an audit. In this way, each audit can be considered a 'reset' of the process.
6. We assume that as time since the last audit passes, the cumulative probability increases that facilities will change state from compliant to non-compliant.
7. Finally, we assume that the only way that facilities can return to the compliant state from the non-compliant state is as a consequence of an (unannounced) audit, that is, there is no spontaneous reversion to compliance. Equivalently, we could say that non-compliance that spontaneously reverts to compliance without an unannounced audit is out of scope for the model being developed.[1]

[1] Reviewer RC pointed out, correctly we feel, that occasionally it is possible for an entity to become non-compliant but then revert to being compliant. An example might be a person going on holiday,

In order to estimate the compliance rate of the AA population under a particular audit regime, we constructed a model that predicts the audit outcome as a function of time since last audit, among other things. We can the specify what the time-since-last audit would look like for the population under the different counterfactual audit regimes, that is, we can say what the time-since-last audit would look like under 6-month audits, 12-month audits, etc. We then used these time-since-last-audit values in the model to predict the compliance status of the population.

4.3. Materials and Methods

4.3.1. Data

The department provided detailed data for a total of 95,243 audits undertaken from August 1999 to September 2023 inclusive. These audits comprised approximately 5 million criteria, with at least 1 and up to about 500 criteria per audit, mode of 32, mean 53.5, and median 43. The audit records included 13,321 in-scope AA's (past and present) with the most-audited AA being audited 76 times, and a mode of 1 audit, minimum 1, mean 7.1, and median 4. These formed the group of AA's from which we sought at least two of the in-scope audits.

Upon reduction to the appropriate scope and agreed timeframe (namely, the last 10 years of audits), the audit records covered a total of 8,050 in-scope AA's, with the most-audited AA being audited 39 times, and a mode of 1 audit, minimum 1, mean 4.7, and median 4.

4.3.2. Complications

As is common with regulatory data, complications abound.

1. Audit outcomes comprise criterion outcomes (also referred to as *conditions*). A simple but not trivial algorithm is used by the department to combine the criterion outcomes into the audit outcome. Under reasonable conceptual models of governance, if at least one criterion is non-compliant then it is more likely that the others will also be non-compliant (see, *e.g.* Lane *et al.*, 2017). Therefore, the criterion assessment outcomes within an audit may or may not be statistically independent of one another, which creates statistical complications in modeling. Hence, it is a material and open question whether the audit outcome or the criterion outcome should be treated as the fundamental observation.

2. The classes for which an AA is approved are considered to influence the likelihood of passing an audit. A range of factors may be relevant: the number of criteria, the type of business enterprise and the operator drawn to that business, the ease or difficulty of complying with the class criteria, etc. Hence, knowing the classes that obtain and to which the audit refers would be valuable information for predicting compliance. However, audits of AA's are held across one or more

the replacement not doing things quite right, but the situation being resolved upon the holiday returning—possibly with no knowledge that such thing has happened. Alternately some other process might be used because of a temporary shortage. However we are not concerned with the potential impact of these edge cases on the generality of our conclusions.

Table 4.1.: Two options for statistical modeling of the AA audit data. The *Unit* column identifies the fundamental unit for analysis, and n the concomitant number of observations. *Complications* reports the shortcomings.

Unit	n	Complications
Audit	22,211	Audits of AA's apply across all AA classes pertaining to the AA. AA class is not recorded during the audit. Non-compliances apply at the AA level, not the class level, so audit outcomes are across all criteria.
Criterion	1,029,025	Audit-level fails are the result of combinations of critical, major, and minor non-compliances at the criterion level (also referred to as *conditions*). Some criteria are optional. Criteria can drop in and out of the class requirement. Edited criteria are assigned a new number. Criterion outcomes may be correlated within audits, and this correlation may vary in time.

classes depending on how many class registrations the AA holds. All of the criteria for all of the approved classes for the AA are assessed in the audit. Each of these criteria may be relevant to more than one class. The classes for which AA's are audited at the time of audit are not possible to infer directly from the available data unless (i) inferred from criteria that were assessed in the audit, which is unwholesomely complicated, or (ii) accessed from cross-referencing the class registry database, under which a substantial number of in-scope audits occurred whilst the AA was not approved for any class.

3. Under the conceptual model outlined in Section 4.2, the outcome of an audit can be thought of as time-bound. That is, the felicity of the audit's outcome as a signifier for the compliance status of the AA reduces as time passes — in a sense, the audit outcomes get old, because the status is more likely to have changed in the intervening time since the audit if the intervening time is longer, all else being equal. If criteria are held to be fundamental then we have to consider the unpalatable possibility that the outcomes for different criteria may have different aging rates.

4. AA's can change class status, hence audits on the AA may be for different sets of classes and therefore different sets of criteria.

5. Detailed criterion outcome capture available for only the past 10 years.

Table 4.1 outlines the distinctions between analyzing the data at each of the two levels (audit, or criterion within audit). In order to resolve and simplify these considerations, we elected to choose the audit as the fundamental observation and to ignore the class identity entirely in order to develop a model that would have regulatory utility. Brief notes are provided in the appendices on models that approach the process differently; audit-level models that include class information can be found in Appendix B and models that assume that the criterion is the fundamental observation can be found in Appendix C.

4.3.3. Data exclusions

We reduced the dataset to the last 10 years of data because records preceding 2013 were subject to inconsistent outcome recording. We excluded audits of international facilities, audits with pending results, and audits with `NULL` status. We also excluded customs broker (class 19) AA's and arrangements relating to proclaimed ports (airports and sea ports), first points of entry (airports and sea ports) international mail centres, and food importer compliance agreements. We included only scheduled or probationary audits that had a preceding scheduled or probationary audit. We included scheduled and probationary audits as these are the only audits that involve compliance assessment of all the conditions of the AA with which the operator is legally obliged to comply. Finally we removed 28 audits that had a between-audit time of less than 1 week.

Figure 4.1 reports the distribution of observations at the conclusion of the data exclusion process. The figure shows that the preponderance of data are available for the announced audits for which the previous audit was a pass, and very much fewer for the unannounced audits for which the previous audit was a fail.

4.3.4. The AA's approved at September 1, 2023

Furthermore, the department provided a set of AA arrangement IDs of those AA's that were approved as at September 1, 2023. A total of 3060 distinct arrangements were approved at that time. This AA count reduced to 2850 when we excluded AA's for which we did not have scheduled or probationary audits within the 10-year time-span that we decided upon. The time since the last scheduled or probationary audit of these AA's is presented in Figure 4.2. These data are used to estimate the number of AA's that would pass an audit if it were taken as at September 1, 2023 reported in Section 4.5.

4.4. Statistical model

The dataset is not merely a sample from a population to which we wish to draw inference. We have access to all of the available audit results. Consequently our goal is not to draw inference about a population from which a sample is collected, but instead to draw inference about a process from which the sample is generated, and make predictions about future instances of that sample.

After considerable experimentation, we settled on Generalised Estimating Equations (GEE) at the audit level as the preferred approach for modeling (Hardin & Hilbe, 2012). The response variable was whether or not the audit of the arrangement identified sufficient non-compliance to record a fail, which is a binary outcome. There was a strong possibility of between-audit within-arrangement correlation, which meant that use of generalised linear models (including logistic regression and probit analysis) would not be appropriate. The rarity of the failure event meant that a conditional model, such as a generalised linear mixed-effects model, would have bias in its predictions (confirmed by experimentation). Since our interest was in making predictions from a marginal model, yet we were concerned that the assumption of within-arrangement independence would be suspect, we elected to use GEE.

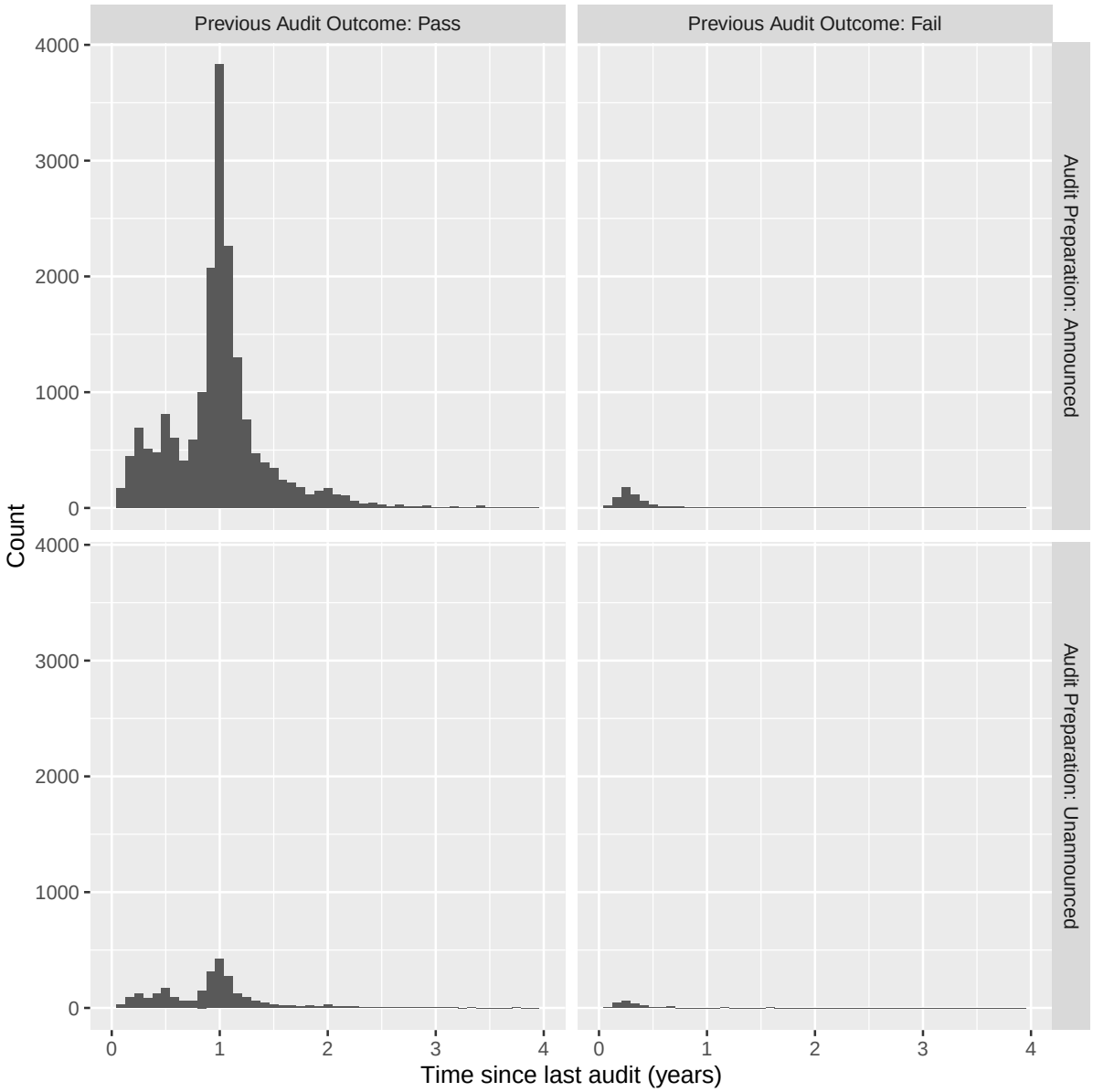

Figure 4.1.: Histograms of the time since last audit by whether the previous audit was sa pass or fail (column) and whether the current audit is announced or unannounced (row).

4.5. Results and discussion

Extensive testing established that the only terms needed in the predictive model were the main effects of (i) a four-knot smooth spline function of duration (meaning, time since last audit), (ii) the outcome of the previous audit, and (iii) the preparation of the current audit (announced or unannounced). Potential interactions were tested and discarded because they were not statistically significant. The relevant statistical model prediction as a function of the relevant predictors is presented in Figure 4.3.

Figure 4.3 reports the predicted probability that an *unannounced* audit of a random AA held at time x years after its previous audit would fail given that the previous audit was a fail (red) or not (green). The solid lines show the best prediction and the transparent ribbons show the approximate 95% confidence intervals around the best

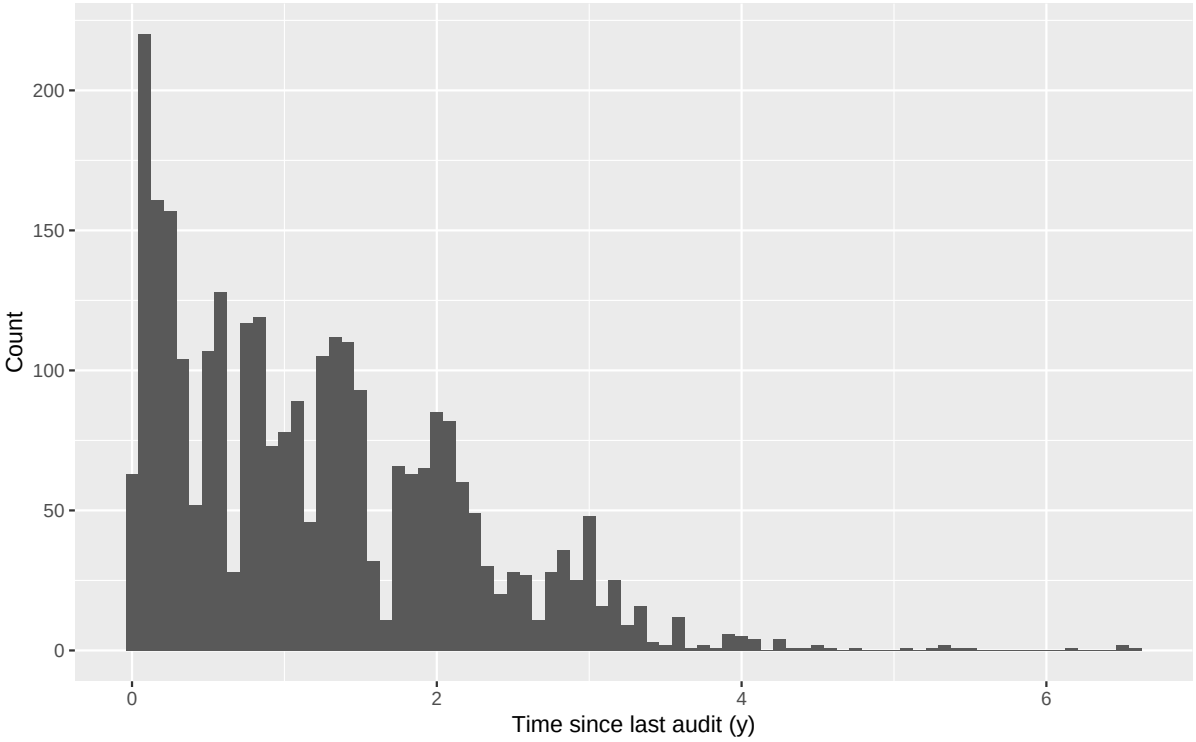

Figure 4.2.: Histogram of the time since the last audit for the approved AA's at September 1, 2023. Each bar represents a month.

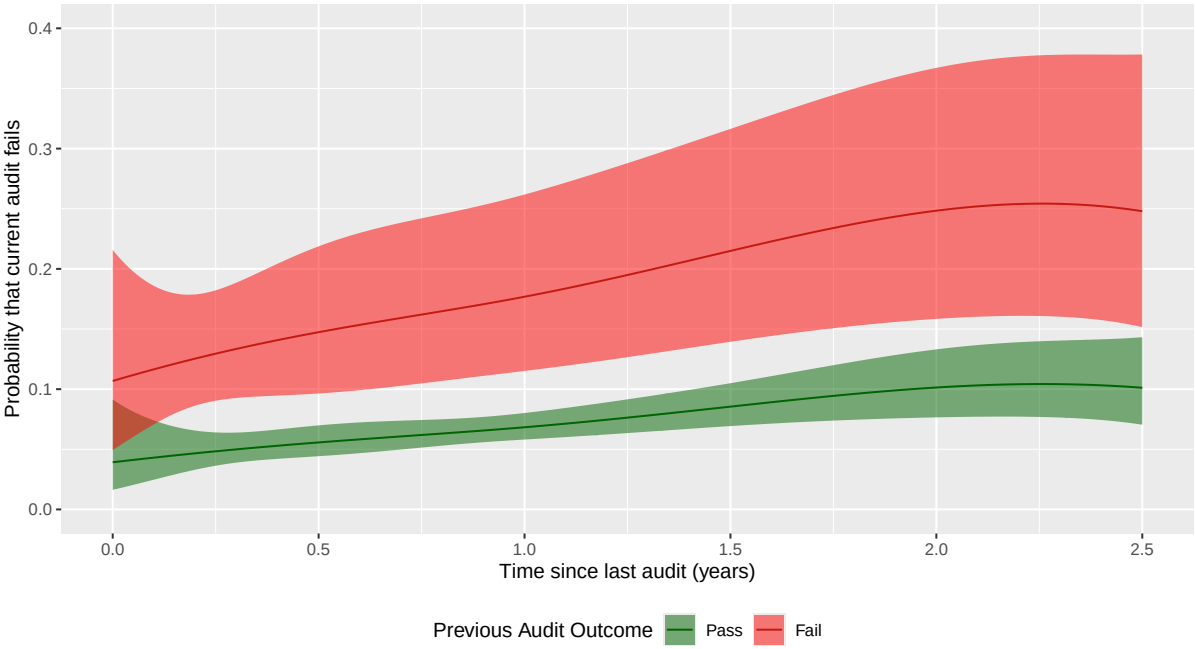

Figure 4.3.: Modelled proportion of unannounced audit-level failures, distinguished by outcome of previous audit. The solid lines show the best prediction and the transparent ribbons show the approximate 95% confidence intervals around the best predictions.

predictions. The odds that an audit will record a fail given that the previous audit for that AA was a fail compared to the situation where the previous audit was not a fail

are about 3:1 (95% ci 2:1–5:1; Section A).

The first notable characteristic is that, as is intuitively satisfying, regardless of how long it has been since the previous audit, the probability of the current audit failing is higher if the previous audit were a fail than if it were a pass. This suggests that there is some kind of history effect, although it could also be a consequence of omitting important predictors (such as class, or the AA identity, potentially). The second notable characteristic is that the trajectory for each line is uniformly upwards from $x = 0$ to about $x = 2.5$ years, after which the data become sparse and the outcome regulatorially uninteresting. We can interpret this as suggesting that a longer time period between audits correlates to a higher probability that the AA's state has switched from compliant to non-compliant, that is, over time the likelihood of non-compliance increases.

The predictions from the statistical model were then applied to the state-space model documented in Section A.1 and used to compute the probability of achieving a given population-level compliance as a function of the time between audits, as presented in Figure 4.4. Three approaches were used, namely: a naive approach, a bootstrap approach and a cohort bootstrap approach (see Section A.1 for further detail). We opted for the latter because it provided the best conceptual match to the process. Consequently the figure has three lines for each of the four candidate values for audit timing. In addition we captured the same information for the current distribution of predicted failure rates, that is, instead of predicting the failure probabilities for exact between-audit times (e.g., 9 months) we predicted the failure probabilities using the observed time since last audit and previous audit outcomes from the data, computed as at September 1, 2023. This (single) trajectory, labeled 'Current' in Figure 4.4 (coloured pink), is almost indistinguishable from the 18-month Cohorts estimate (this is the green dotted line).

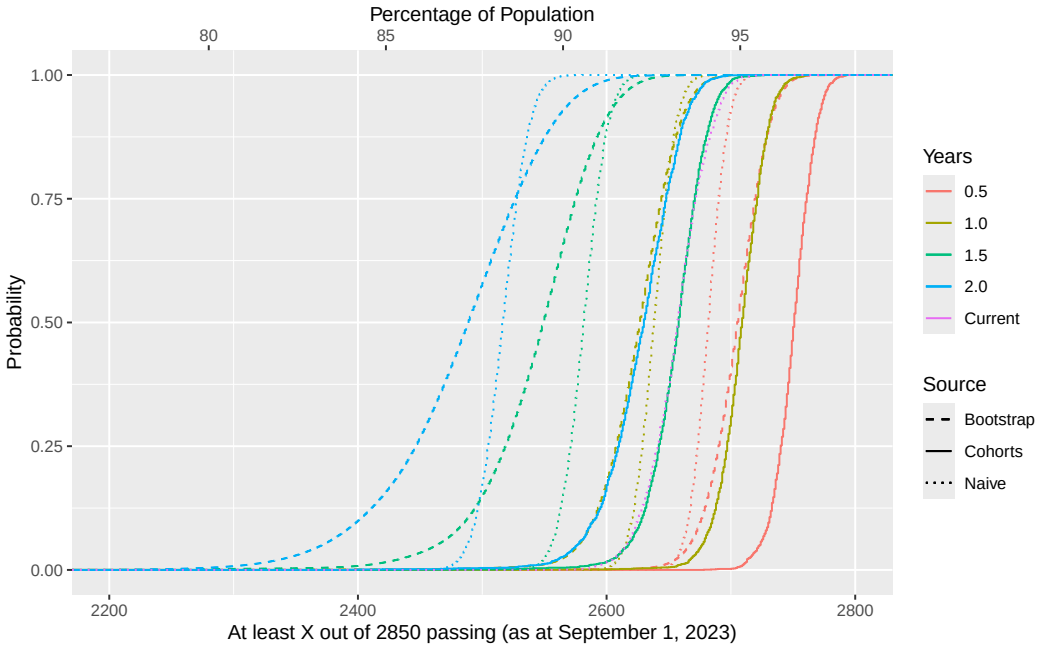

Figure 4.4.: The lower boundary on the simulated number of AA's that would pass an unannounced audit with given probability as a function of the time between audits, from the 2850 AA's approved at September 1, 2023.

Figure 4.5 shows the long-run modelled relationship between audit rate and compliance level, contextualised by probability. Each line colour represents a different quantile (level) of probability. For example, considering the set of 2850 AA's that were approved at September 1, 2023, there is an estimated 95% probability that at least 2575 AA's would pass an unannounced audit if the audits were every 24 months; this increases by 100 to 2675 if the audits were every 12 months. The right axis is the proportion of the population, and the top axis reports the number of audits per year implied by the audit period. Hence we can say, for example, there is an estimated 0.95 probability that at least 94% of the AA's will be compliant under a audit period of 12 months, which corresponds to 2850 audits per year, or reframing the observation in the context of the motivating question, that an audit rate of just 2850 per year leads to a compliance level of at least 94% with estimated probability 0.95.

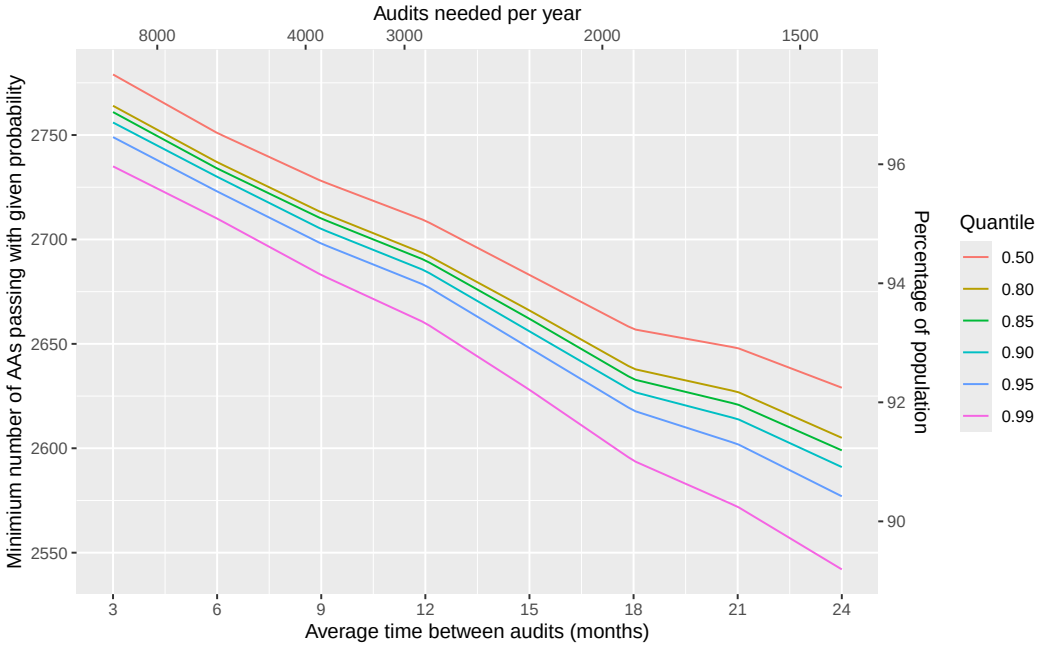

Figure 4.5.: The simulated minimum expected number of AA's that would pass an unannounced audit with given probability from the 2850 AA's approved at September 1, 2023.

As noted in the appendix that provides the modeling details (Appendix A), the quantiles presented in Figure 4.5 include the key elements of uncertainty such as parameter uncertainty.

Finally, as noted above, the odds that an audit will fail are higher if the previous audit for the arrangement failed. This result suggests that reducing the time between audits for the cohort that failed their previous audit will improve compliance overall. That is, more frequent audits of the failing cohort coupled with less frequent audits of the passing cohort may result in the same population-level compliance with a reduction in effort. This approach is analogous to the department's Compliance-Based Intervention Scheme.[2]

[2]CBIS: https://www.agriculture.gov.au/biosecurity-trade/import/goods/plant-products/risk-return

5. Conclusion

This report details a modelling framework to help choose an idealised time-between-audits based on the likely impact on the probable number of AA's that are compliant. Our results suggest that the current audit regime has characteristics that are similar to an 18-month between-audit period (see Figure A.3). This corresponds to approximately 1900 audits a year, on average, leading to a compliance level of just over 92% (2625/2850) with estimated probability 0.9 (see Figure 1). The model suggests that if the inter-audit period was reduced to be approximately 12 months, corresponding to 2850 audits per year, then the system estimates a compliance level of just over 94% with estimated probability 0.9. This outcome provides a useful comparison for guiding choices about the time between audits. Furthermore, the model suggests that increasing the audit rate on AA's for which the previous audit was a fail may efficiently increase the population level compliance, although formal exploration of this possibility was not within the scope of the current project.

The outcome of the report is built upon a coupling of a statistical model, fitted by GEE, and a mathematical model (a Markov chain. In a sense, the importance of being able to connect these two models provided some natural constraints on the ideal complexity of each. A more complex model could plausibly yield more precise predictions but would not be so easy to deploy or to understand. We chose models that represented a reasonable trade-off between prediction quality (statistical fastidiousness) and simplicity (operational relevance).

A reviewer[1] suggested that it might also make sense to consider "arrangement-days of compliance" as an objective function (as opposed to probability of compliance), which is an interesting wrinkle. Under such an approach, the aim would be to detect non-compliance as quickly as possible rather than to achieve a specific level of compliance (Cannon, 2009). However, under our model, in which audits effectively reset the compliance status of the audited arrangements, it may make no practical difference compared with our current approach.

It is worth noting that the results of the modeling exercise documented in Section 4.5 show a considerable effect of unannounced as opposed to announced audits, with the odds of an unannounced audit detecting a failing state being about three times higher than the odds of an announced audit detecting a failing state. This finding possibly has implications for any policy around whether audits should be announced.

Finally, very few of the assumptions or modeling choices are contingent on specific details of biosecurity risk management, so the overall approach should generalise easily to other audit settings.

[1] RC

Bibliography

Advani A, Elming W, Shaw J (2023) The Dynamic Effects of Tax Audits. *The Review of Economics and Statistics*, **105**, 545–561. URL https://doi.org/10.1162/rest_a_01101.

Allingham MG, Sandmo A (1972) Income tax evasion: a theoretical analysis. *Journal of Public Economics*, **1**, 323–338. URL https://www.sciencedirect.com/science/article/pii/0047272772900102.

Anton WRQ, Deltas G, Khanna M (2004) Incentives for environmental self-regulation and implications for environmental performance. *Journal of Environmental Economics and Management*, **48**, 632–654. doi:10.1016/j.jeem.2003.06.003. URL https://www.sciencedirect.com/science/article/pii/S0095069603001025.

Avenhaus R, Canty MJ (2009) Inspection Games. In: *Encyclopedia of Complexity and Systems Science* (ed. Meyers RA), pp. 4855–4868. Springer, New York, NY. URL https://doi.org/10.1007/978-0-387-30440-3_287.

Avenhaus R, Von Stengel B, Zamir S (2002) Inspection games. In: *Handbook of Game Theory with Economic Applications*, vol. 3, pp. 1947–1987. Elsevier. URL https://www.sciencedirect.com/science/article/pii/S157400050203014X.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48. doi:10.18637/jss.v067.i01.

Becker GS (1968) Crime and Punishment: An Economic Approach. *Journal of Political Economy*, **76**, 169–217. URL https://www.jstor.org/stable/1830482.

Bloomquist K (2011) Tax Compliance as an Evolutionary Coordination Game: An Agent-Based Approach. *Public Finance Review*, **39**, 25–49. URL https://doi.org/10.1177/1091142110381640.

Cannon RM (2009) Inspecting and monitoring on a restricted budget—where best to look? *Preventive veterinary medicine*, **92**, 163 – 174. URL http://linkinghub.elsevier.com/retrieve/pii/S0167587709001822.

CSIRO (2020) Australia's Biosecurity Future: Unlocking the next decade of resilience. Tech. rep., Commonwealth Scientific and Industrial Research Organisation.

DAFF (2023) Approved arrangements general policies, Version 7.3. URL https://www.agriculture.gov.au/biosecurity-trade/import/arrival/arrangements/general-policies.

Das S, Geedipally SR, Dixon K, Sun X, Ma C (2019) Measuring the Effectiveness of Vehicle Inspection Regulations in Different States of the U.S. *Transportation Research Record*, **2673**, 208–219. URL https://doi.org/10.1177/0361198119841563.

DeBacker J, Heim BT, Tran A, Yuskavage A (2018) Once Bitten, Twice Shy? The Lasting Impact of Enforcement on Tax Compliance. *The Journal of Law and Economics*, **61**, 1–35. URL https://www.journals.uchicago.edu/doi/abs/10.1086/697683.

degl'Innocenti DG, Levaggi R, Menoncin F (2022) Tax avoidance and evasion in a dynamic setting. *Journal of Economic Behavior & Organization*, **204**, 443–456. URL https://www.sciencedirect.com/science/article/pii/S0167268122003857.

Dodin B, Elimam A, Rolland E (1998) Tabu search in audit scheduling. *Eur. J. Oper.*

Res., **106**, 373–392. doi:10.1016/S0377-2217(97)00280-4.

Duflo E, Greenstone M, Pande R, Ryan N (2018) The Value of Regulatory Discretion: Estimates From Environmental Inspections in India. *Econometrica*, **86**, 2123–2160. Eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA12876.

Earnhart D (2004) Regulatory factors shaping environmental performance at publicly-owned treatment plants. *Journal of Environmental Economics and Management*, **48**, 655–681. URL https://www.sciencedirect.com/science/article/pii/S0095069603001347.

Earnhart D, Glicksman RL (2015) Extent of Cooperative Enforcement: Effect of the Regulator-Regulated Facility Relationship on Audit Frequency. *Strategic Behavior and the Environment*, **5**. URL https://papers.ssrn.com/abstract=2676856.

Earnhart D, Harrington DR (2021) Effects of audit frequency, audit quality, and facility age on environmental compliance. *Appl. Econ.*, **53**, 3234–3252. doi:10.1080/00036846.2020.1854449.

Earnhart D, Leonard JM (2013) Determinants of environmental audit frequency: The role of firm organizational structure. *Journal of Environmental Management*, **128**, 497–513. URL https://www.sciencedirect.com/science/article/pii/S030147971300371X.

Eckert H (2004) Inspections, warnings, and compliance: the case of petroleum storage regulation. *Journal of Environmental Economics and Management*, **47**, 232–259. URL https://www.sciencedirect.com/science/article/pii/S0095069603000792.

Evans MF, Liu L, Stafford SL (2011) Do environmental audits improve long-term compliance? Evidence from manufacturing facilities in Michigan. *Journal of Regulatory Economics*, **40**, 279–302. URL https://doi.org/10.1007/s11149-011-9163-2.

Gemmell N, Ratto M (2012) Behavioral responses to taxpayer audits: evidence from random taxpayer inquiries. *National Tax Journal*, **65**, 33–57. URL https://www.journals.uchicago.edu/doi/abs/10.17310/ntj.2012.1.02.

Gray WB, Deily ME (1996) Compliance and Enforcement: Air Pollution Regulation in the U.S. Steel Industry. *Journal of Environmental Economics and Management*, **31**, 96–111. URL https://www.sciencedirect.com/science/article/pii/S0095069696900340.

Gray WB, Mendeloff JM (2005) The Declining Effects of OSHA Inspections on Manufacturing Injuries, 1979–1998. *ILR Review*, **58**, 571–587. URL https://doi.org/10.1177/001979390505800403.

Gray WB, Shimshack JP (2011) The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence. *Review of Environmental Economics and Policy*, **5**, 3–24. URL https://www.journals.uchicago.edu/doi/abs/10.1093/reep/req017.

Hanna RN, Oliva P (2010) The Impact of Inspections on Plant-Level Air Emissions. *The B.E. Journal of Economic Analysis & Policy*, **10**. URL https://www.degruyter.com/document/doi/10.2202/1935-1682.1971/html.

Hardin JW, Hilbe JM (2012) *Generalized Estimating Equations*. Chapman & Hall, 2nd edn.

Harrington W (1988) Enforcement leverage when penalties are restricted. *Journal of Public Economics*, **37**, 29–53. URL https://www.sciencedirect.com/science/article/pii/0047272788900035.

Hashimzade N, Myles GD, Page F, Rablen MD (2014) Social networks and occupational choice: The endogenous formation of attitudes and beliefs about tax compliance. *Journal of Economic Psychology*, **40**, 134–146. doi:10.1016/j.joep.2012.09.002.

Hashimzade N, Myles GD, Tran-Nam B (2013) Applications of Behavioural Economics to Tax Evasion. *Journal of Economic Surveys*, **27**, 941–977. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6419.2012.00733.x.

Humblet MF, Saegerman C (2023) Internal audits as a tool to assess the compliance with biosecurity rules in a veterinary faculty. *Front. Vet. Sci.*, **10**. doi:10.3389/fvets.2023.960051.

Hut-Mossel L, Ahaus K, Welker G, Gans R (2021) Understanding how and why audits work in improving the quality of hospital care: A systematic realist review. *PLoS ONE*, **16**, e0248677. doi:10.1371/journal.pone.0248677.

Ivers N, Jamtvedt G, Flottorp S, *et al.* (2012) Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews*, p. CD000259.

Kasper M, Alm J (2022) Audits, audit effectiveness, and post-audit tax compliance. *Journal of Economic Behavior & Organization*, **195**, 87–102. URL https://www.sciencedirect.com/science/article/pii/S0167268122000099.

Kasper M, Rablen MD (2023) Tax compliance after an audit: Higher or lower? *Journal of Economic Behavior & Organization*, **207**, 157–171. URL https://www.sciencedirect.com/science/article/pii/S0167268123000136.

Lane SE, Arthur AD, Aston C, Zhao S, Robinson AP (2017) When does poor governance presage biosecurity risk? *Risk Analysis*, **38**, 653–665. URL https://doi.org/10.1111%2Frisa.12873.

Laplante B, Rilstone P (1996) Environmental Inspections and Emissions of the Pulp and Paper Industry in Quebec. *Journal of Environmental Economics and Management*, **31**, 19–36. URL https://www.sciencedirect.com/science/article/pii/S0095069696900297.

Maciejovsky B, Kirchler E, Schwarzenberger H (2007) Misperception of chance and loss repair: On the dynamics of tax compliance. *Journal of Economic Psychology*, **28**, 678–691. URL https://www.sciencedirect.com/science/article/pii/S0167487007000128.

Medu O, Turner H, Cushon JA, *et al.* (2016) Restaurant inspection frequency: The RestoFreq Study. *Canadian Journal of Public Health*, **107**, e533–e537. URL https://doi.org/10.17269/CJPH.107.5399.

Mittone L, Panebianco F, Santoro A (2017) The bomb-crater effect of tax audits: Beyond the misperception of chance. *Journal of Economic Psychology*, **61**, 225–243. URL https://www.sciencedirect.com/science/article/pii/S0167487016306341.

Newbold KB, McKeary M, Hart R, Hall R (2008) Restaurant inspection frequency and food safety compliance. *Journal of Environmental Health*, **71**, 56–61.

Porter WP, Horn MJ, Cooper DM, Klein HJ (2013) Auditing laboratory rodent biosecurity programs. *Lab Animal*, **42**, 427–431. doi:10.1038/laban.409.

Racicot M, Venne D, Durivage A, Vaillancourt JP (2012) Evaluation of strategies to enhance biosecurity compliance on poultry farms in Quebec: Effect of audits and cameras. *Preventative Vetrinary Medicine*, **103**, 208–218. doi:10.1016/j.prevetmed.2011.08.004.

Roberts F, Amirkhanyan A, Meier KJ, Davis J (2022) Limiting managerial discretion by regulation: nursing homes and the national background check program. *International Public Management Journal*, **25**, 392–412. URL https://doi.org/10.1080/10967494.2022.2036884.

Rossi R, Tarim SA, Hnich B, Prestwich S, Karacaer S (2010) Scheduling internal audit activities: a stochastic combinatorial optimization problem. *J. Comb. Optim.*, **19**, 325–346. doi:10.1007/s10878-009-9207-z.

Rossiter A, Hester SM (2017) Designing Biosecurity Inspection Regimes to Account for Stakeholder Incentives: An Inspection Game Approach. *Economic Record*, **93**, 277–301. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-4932.12315.

Rönninger S, Holmes M (2009) A risk-based approach to scheduling audits. *PDA journal of pharmaceutical science and technology*, **63**, 575–88.

Sandberg M, Dahl J, Lindegaard LL, Pedersen JR (2017) Compliance/non-compliance with biosecurity rules specified in the Danish Quality Assurance system (KIK) and Campylobacter-positive broiler flocks 2012 and 2013. *Poultry Science*, **96**, 184–191. doi:10.3382/ps/pew277.

Saw ST (2017) Does frequency of audits improve taxpayer compliance? *South East Asia Journal of Contemporary Business, Economics and Law*, **14**, 18–26.

Shapiro D, Stewart-Brown B (2008) Farm biosecurity risk assessment and audits. In: *Avian influenza* (ed. Swayne DE), p. 369. Blackwell Publishing, 1 edn.

Stroup WW, Ptukhina M, Garai J (2024) *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Taylor & Francis, 2nd edn.

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

Telle K (2004) Effects of inspections on plants' regulatory and environmental performance - evidence from Norwegian manufacturing industries. Working Paper 381, Discussion Papers, Oslo Norway. URL https://www.econstor.eu/handle/10419/192363.

Telle K (2009) The threat of regulatory environmental inspection: impact on plant performance. *Journal of Regulatory Economics*, **35**, 154–178. URL https://doi.org/10.1007/s11149-008-9074-z.

Telle K (2013) Monitoring and enforcement of environmental regulations: Lessons from a natural field experiment in Norway. *Journal of Public Economics*, **99**, 24–34. URL https://www.sciencedirect.com/science/article/pii/S0047272713000145.

Traxler C (2014) Deterrence of Tax Evasion. In: *Encyclopedia of Criminology and Criminal Justice* (eds. Bruinsma G, Weisburd D), pp. 1005–1014. Springer, New York, NY. URL https://doi.org/10.1007/978-1-4614-5690-2_411.

Vanegas L, Rondón L, Paula G (2023) Generalized estimating equations using the new r package glmtoolbox. *The R Journal*, **15**, 105–133. doi:10.32614/RJ-2023-056. Https://doi.org/10.32614/RJ-2023-056.

Vanegas LH, Rond'on LM, Paula GA (2024) *glmtoolbox: Set of Tools to Data Analysis using Generalized Linear Models*. URL https://CRAN.R-project.org/package=glmtoolbox. R package version 0.1.11.

Wood S (2017) *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edn.

Zarembski AM, Attoh-Okine N, Boyce TJ (2017) Risk-based scheduling methodology for audit inspections of curves on high-speed mainline tracks. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, **232**, 1650–1659. doi:10.1177/0954409717740748.

Revision History

Revision	Date	Author(s)	Description
1.0.0	April 18 2024	NM	Copied from Report 1
1.1.0	May 1 2024	NM	Enhanced literature review
2.0.0	August 11 2024	AR	Updated modeling
2.1.0	August 30 2024	AR	Developed relationship between compliance probability and audit return rate
3.0.0	October 2 2024	AR	Feature-complete draft of main body of report provided to NS and RW for feedback
3.1.0	October 9 2024	NS, RW	Updated following review comments
4.0.0	October 20 2024	AR	Draft for review, with analysis documented, including brief coverage of unfruitful directions in appendices.
5.0.0	December 17 2024	AR	Final draft following SRP review; no changes required from DAFF review.

A. Technical details for the adopted approach

The goal of the modeling exercise is to develop a statistical model that would enable a statement about the proportion or number of AA's that are failing with probability x as a consequence of the design of the audit regime. We interpret this as: "assuming that the AA verification were being undertaken by using audit regime R, if all the 2850 AA's were audited immediately unannounced, then at least $f\%$ of the audit outcomes would be fails with probability p." Here we focus on four candidate audit regimes which differ only in the time to return, in other words, the set period between audits, namely 6 months, 12 months, 18 months, and 24 months.

A.1. Models

This question requires a statistical model that predicts the the outcome of an audit of an AA as a function of various explanatory factors, for example, the classes held by the AA being audited, the time since the last audit, the outcome of the last audit, and so on. As noted in the report, we opted for a Generalised Estimating Equations (GEE) fitting approach after considerable experimentation. Generalised linear model fitting approaches such as logistic regression and probit analysis were not suitable because they require the assumption of conditional independence between the observations, and we considered it likely that observations within AA's would not be conditionally independent. In other studies we have used generalised linear mixed-effects models (glmm's) to manage such issues but in this case our interest is in obtaining population average predictions and the glmm predictions are conditional on the random effects, whereas we want to make predictions that are marginal.

We used the glmtoolbox package of R (Vanegas *et al.*, 2023, 2024), because it enabled the construction of prediction intervals (other GEE-supporting packages did not, to the best of our knowledge). Part of the GEE model infrastructure is specification of a subject, within which the observations may be conditionally dependent; we used the AA identifier to identify audits of a common arrangement. Furthermore it is necessary to choose a correlation structure (against the specification of which the GEE model is, fortunately, robust); we opted for an *exchangeable* structure, which sets out that all pairs of observations (audits) within a cluster (arrangement) have the same correlation, regardless of how far apart they are in time. Other supported options were ruled out because the data were insufficient, for example, temporally-sensitive versions require the data to be presented on an equi-spaced grid (Vanegas *et al.*, 2023), the unstructured version resulted in estimates of the correlation matrix that were not positive definite, and the independent version was a poorer fit according to the tests documented in (Vanegas *et al.*, 2023).

Accepting the exchangeable correlation structure, we fit a large number of nested

models, and experimented with different ways of handling the duration (time since last audit), opting in the end for a 4-knot B-spline[1]. After this elimination we included only the following terms: the 4-knot B-spline function of time since last audit (`bs(duration, 4)`), whether the current audit is announced or unannounced (`preparation`), and whether the previous audit was a pass or a fail (`previous.fail`). The following R output reports a set of nested whole-model comparisons that shows that, among the predictors that made the final cut, only the main effects were statistically worth retaining.

```
Wald test

Model 1 :  fail ~ 1
Model 2 :  fail ~ bs(duration, 4)
Model 3 :  fail ~ bs(duration, 4) + audit.preparation
Model 4 :  fail ~ bs(duration, 4) + audit.preparation +
                  previous.fail
Model 5 :  fail ~ bs(duration, 4) + audit.preparation +
                  previous.fail + bs(duration, 4):audit.preparation
Model 6 :  fail ~ bs(duration, 4) + audit.preparation +
                  previous.fail + bs(duration, 4):audit.preparation +
                  bs(duration, 4):previous.fail

          Chi    df   Pr(>Chi)
1 vs 2  15.7883   4   0.003317 **
2 vs 3 159.3284   1  < 2.2e-16 ***
3 vs 4  21.3820   1  3.763e-06 ***
4 vs 5   9.8829   4   0.042447 *
5 vs 6   3.4590   4   0.484133
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Models 5 and 6 allow for the effect upon the audit outcome of the passing of time since the previous audit to vary depending on whether the current audit is announced or unannounced (not significant) and whether the previous audit was a fail (again, not significant). The parameter estimates and summary information for the model are presented below.

```
Sample size
   Number of observations:  21968
       Number of clusters:  4620
                            Min  25%  50%  75%  Max
            Cluster sizes:    1    2    5    7   14
************************************************************
Model
        Variance function:  binomial
            Link function:  logit
    Correlation structure:  Exchangeable
************************************************************
Coefficients
```

[1] https://en.wikipedia.org/wiki/B-spline

	Estimate	Std.Error	z-value	Pr(>|z|)
(Intercept)	-4.42477	0.41503	-10.66128	< 2.22e-16
bs(duration, 4)1	0.43227	0.63019	0.68593	0.49276
bs(duration, 4)2	0.68213	0.48006	1.42092	0.15534
bs(duration, 4)3	2.03429	1.30793	1.55535	0.11986
bs(duration, 4)4	-1.07211	2.59365	-0.41336	0.67934
audit.preparation = Unannounced	1.17842	0.09269	12.71365	< 2.22e-16
previous.fail = TRUE	1.09540	0.23689	4.62407	3.7629e-06

The coefficients for the 4-knot B-spline do not bear intuitive interpretation; the output is provided to focus on preparation and previous fail. These results show that the magnitude of the effects of unannounced audits (Odds ratio 3.25; 95% CI 2.70–3.91) and the previous audit being a fail (OR 2.99; 95% CI 1.86–4.80) are about the same, and correspond to an overall duration effect of about 2 years (loosely speaking, from casual inspection of the predicted values that underpin Figure A.1).

The final model is therefore:

$$y_{ij} \overset{d}{=} Be(p_{ij}) \tag{A.1}$$

$$logit(p_{ij}) = f_4(t_{ij}) + o_{ij} + a_{ij} \tag{A.2}$$

where y_{ij} is the outcome of the $j^{th}(j = 2 \dots n_i)$ audit of AA i, p_{ij} is the probability that the audit will fail, $logit$ is the function $logit(x) = log(x/(1-x))$, Be signifies the Bernoulli distribution, f_k signifies a b-spline with 4 knots, t_{ij} is the time between the j^{th} audit and its predecessor, o_{ij} is whether the outcome of the preceding audit was a pass or a fail, and a_{ij} is whether audit j is announced or unannounced.

The predictions of the model are presented in Figure A.1. This reports the predicted probability that an announced or an unannounced audit of a random AA held at time x years after its previous audit would fail given that the previous audit was a fail (blue) or not (red). The solid lines show the best prediction and the transparent ribbons show the approximate 95% confidence intervals around the best predictions.

The first notable characteristic is that, as is intuitively satisfying, regardless of how long it has been since the previous audit, (i) unannounced audits fail at a higher rate than announced audits, and (ii) the probability of the current audit failing is higher if the previous audit was a fail than if it was a pass. This latter observation suggests that there is some kind of history effect, although it could also be a consequence of omitting important predictors (such as class, or the AA identity, potentially). The second notable characteristic is that the trajectory for each line is uniformly upwards from $x = 0$ to about $x = 2.5$ years, after which the data become sparse (the ribbons become trumpets) and the outcome regulatorially uninteresting. We can interpret this as suggesting that a longer time period between audits correlates to a higher probability that failure will be detected, which in turn suggests a higher probability that the latent state has switched from compliant to non-compliant.

A.2. Model connection

Having settled on a model to predict the probability that an AA would fail an audit as a function of the time since the last audit, and other characteristics (namely, whether the audit would be *announced* or *unannounced*, and whether the previous audit was a *pass*

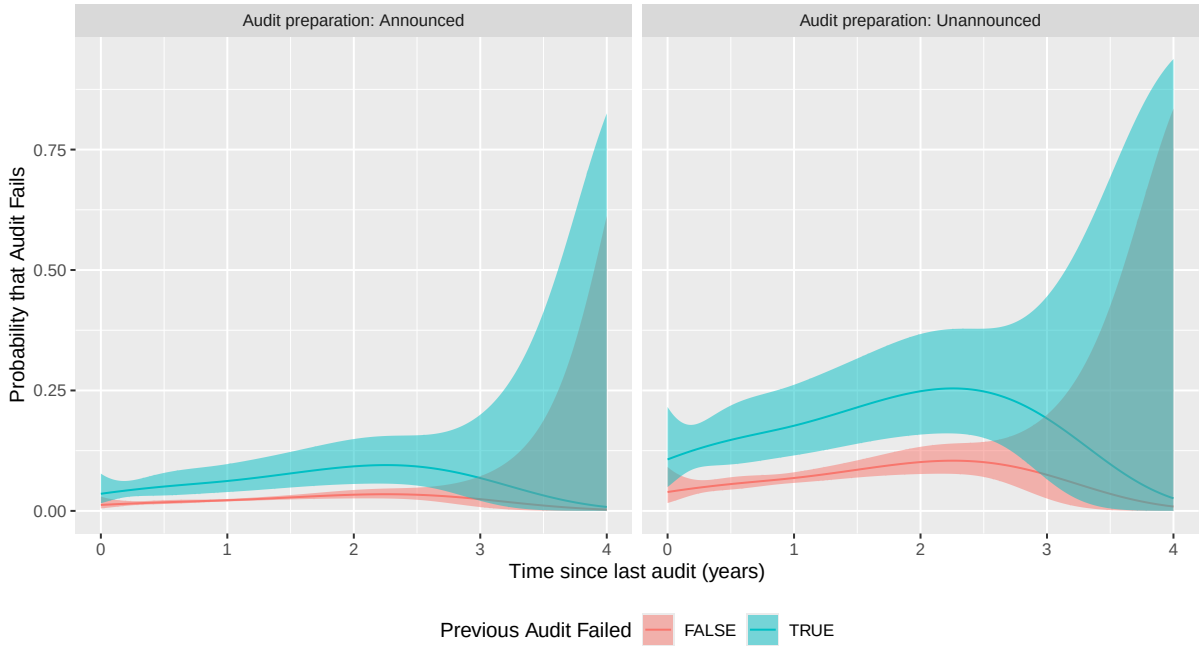

Figure A.1.: The best-supported model to predict audit outcome from time since previous audit, outcome of previous audit, and whether the current audit is announced or not.

or *fail*), we needed to develop an approach to translating the model predictions into an estimate of the proportion of the AA population that would fail an unannounced audit if were undertaken immediately.

We applied the following approach.

1. We assume that an AA $a_i, i = 1 \ldots n$ occupies one of two states, namely *passing* or *failing*, which are distinguished by what would be the outcome of an *unannounced* audit of the AA if it were undertaken immediately.
2. We assume that the AA begins its audit cycle as a *passing* AA, and then with a certain probability it switches (or does not switch) class to be a *failing* AA, which state is then detected at the subsequent audit.
3. We assume that the switching or not of the AA's is independent of one another, and is governed by the model developed in Section A.1.
4. We assume that failing AA's do not change to passing AA's of their own accord, but only do so at the time of the audit, and at that time do so with 100% surety.
5. We assume that audits are carried out with 100% sensitivity (all failing states are detected) and specificity (non-failing states are never mistaken for failing states).

Under these assumptions, the observed compliance status of an AA can be represented by a conditional discrete-time two-state Markov chain model (e.g., Figure A.2), where the conditioning is on t, the time between audits.

Under reasonable conditions,[2] we can write the long run (stationary) probability of state occupancy as a function of the time t since the previous audit as

[2]Namely, that the Markov chain is time-homogeneous, irreducible, and aperiodic, all of which are reasonable here.

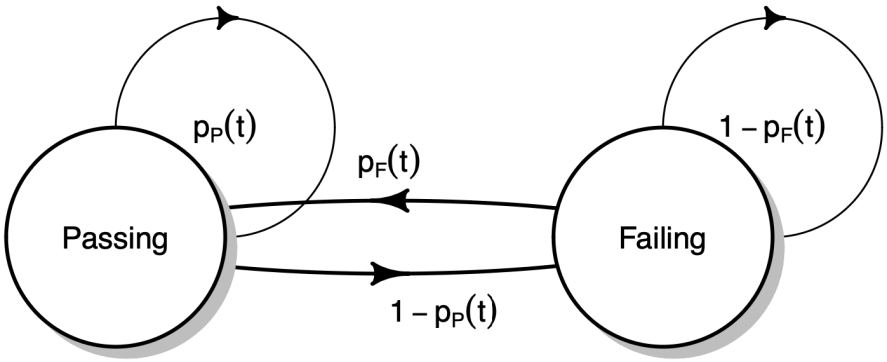

Figure A.2.: A two-state Markov chain model that represents the compliance state of an AA, which can be either *Passing* or *Failing*. $p_P(t)$ is the probability of passing at time t since the previous audit if the previous audit was a pass; $p_F(t)$ if the previous audit was a fail.

$$\pi_P(t) = \frac{p_F(t)}{(1 - p_P(t)) + p_F(t)} \tag{A.3}$$

where $p_P(t)$ and $p_F(t)$ can be predicted directly from the model represented in Figure A.1. We can then make a probabilistic claim about the effect of t upon $\pi_P(t)$ by, for example, plotting the cumulative density function of the Binomial distribution Bi(2850, $\pi_P(t)$). We call this the *naive* estimate.

However, this naive estimate ignores the fact that the predictions from the model are random variables, and that even if the model is true the predictions will vary, and we don't know that the model is true. Therefore probabilistic statements about the effect of t on the compliance rate will be wrong. In order to correct for this gap we used a non-parametric bootstrap, as follows: we carried out the following steps 2000 times,

1. Take a stratified random sample with replacement of size 22,211 from the 22,211 audit outcome observations — this is called a *bootstrap sample*. The strata were formed by the two factors in the model, audit preparation and outcome of previous audit in order to appropriately preserve the relative uncertainties of each combination
2. Refit the final GEE model to the bootstrap sample
3. Predict the audit failure probability at a grid of candidate time to next audit values (we used $t = 6, 12, 18,$ and 24 months assuming that the audits are unannounced)
4. Compute estimates of $\pi_P(t)$, say $\tilde{\pi}_P(t)$, for each time to next audit value
5. Generate 10000 random Binomial observations Bi(2850, $\tilde{\pi}_P(t)$) for each t.

We presented the 2 million observations per value of t as an empirical cumulative density function. We call this the *bootstrap solution*.

Finally, we note that the bootstrap solution is conservative because it evaluates all of the unannounced audit failure probabilities on the basis that the audit will occur at time t after the previous audits, but it ignores the fact that the audits have to be interspersed throughout the year. Therefore, for example for $t = 6$, 437.5 of the 2850

AA's will have been audited less than a month ago, 437.5 audited two months ago, and so on. Consequently, evaluating the risk as though all the AA's were audited 6 months ago, as the state space model assumes, is unnecessarily conservative.

We corrected for this cohort effect by repeating the bootstrap and evaluating the candidate time to next audit values for each month, that is, $t = 1 \ldots 24$, and evaluating the probability at each of the four values for t as an aggregation of the monthly cohorts leading up to each one, assuming uniform distribution of audits per month.

Finally, we compared these results with the compliance rate that is implied by the current distribution of between-audit durations, as presented in Figure 4.2.

A.3. Results

The output of the model outlined in the previous section is presented in Figure A.3. The graph combines the outputs of the various approaches to ease comparison. Five panels are presented, representing the four candidate between-audit durations (namely, 6, 12, 18, and 24 months) and the current distribution (Figure 4.2).

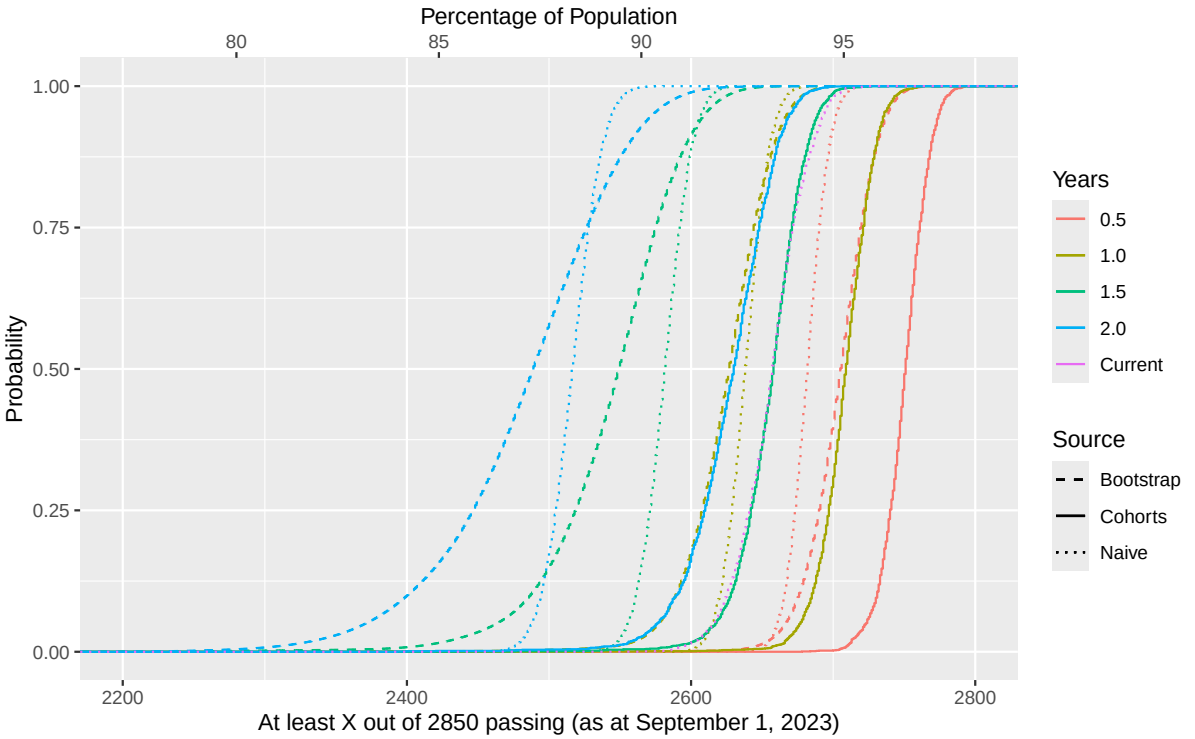

Figure A.3.: The lower boundary on the simulated number of AA's that would pass an unannounced audit with given probability as a function of the time between audits, from the 2850 AA's approved at September 1, 2023.

The plot shows a predictable progression that indicates that as the time between audits increases, the probability distribution of the number of non-compliant arrangements shifts to the right, that is, the model predicts that more arrangements will be non-compliant.

The naive estimate (dashed line) ignores model-based uncertainty (which is expressed as the difference between the ribbons and the lines in Figure A.1), so it uses the best

predictions from Figure A.1 in Equation A.3 to establish the compliance probability distribution.

The bootstrap (solid line) results wrap the naive approach in a non-parametric bootstrap, so, refits the model presented in Figure A.1 to random, with-replacement resamples of the audit data, and for each refit model, plugs the best predictions into Equation A.3, generates random Bernoulli-distributed values, and agglomerates them into a distribution. The blue curve is flatter than the green curve, showing the effect of the greater uncertainty. However, we also see that the bootstrap is correcting for bias in the naive estimates (note that the GEE estimates may not be biased but they are random variables, so curvilinear functions applied to them, such as Equation A.3, can show bias). This is most obvious in the "Return: 0.5" panel, in which the uncertain curve is completely to the left of the naive curve, showing that the naive curve non-compliance rate was probably too high.

Finally, noting that the bootstrap estimate was likely too conservative, we developed the bootstrap cohort estimate (dashed line), which we hold to be the best representation of the process, and shows lower non-compliance rates than the preceding methods.

We used this bootstrap cohort estimate to create Figure 1 to enable comparison of the effect of different times between audits upon the probability of compliance.

B. Class-specific analyses

Here we report an analysis approach that we considered to be a dead end: class-specific audit outcomes. We assessed statistical models of audit-level outcomes with arrangement class as a candidate predictor. It seemed reasonable to us that the outcomes of the audits would be affected by various class-sensitive factors, such as the number of criteria in a class, the number of classes for which an AA was approved at the time of audit, the complexity of the class specifications, the burden of the class requirements, and so on. The analyses are briefly reported here for future consideration.

B.1. Background

These class-sensitive analyses were beset by several complications.

1. Each audit outcome covers all of the classes for which an AA was approved at the time of the audit. Therefore an audit pass may be assumed to apply to all approved classes for the AA, but an audit fail has an unknown provenance, because the outcome is not tied to a particular class (indeed, the class information is not included in the audit data). The only way to deliver a crisp class-specific outcome would be to examine the criteria, the analysis of which is documented in Appendix C, and map those outcomes back to the originating class.
2. The audit outcome algorithm (*e.g.*, a fail results from three or more criteria recording major non-compliances, among other rules) is applied across all the criteria for all the classes, so it is possible that the audit-level fail could be triggered whereas, if they were considered separately, the individual classes would not trigger a fail.
3. Statistical modeling of such data can be a complex undertaking. If the audit-level outcomes are taken to be the observations, and one or more classes may apply at any given observation, and the intent is to allow the model to express the variability of the classes, but the class identity of an AA can change (and indeed is not known but must be inferred from separate records) then constructing the design matrix (a statistical device that is a core component of statistical modeling) must proceed manually.

B.2. Data

The data are as described in Section 3.2. We provide a snapshot of the inferred class-level number of audits and number of fails in Figure B.1.

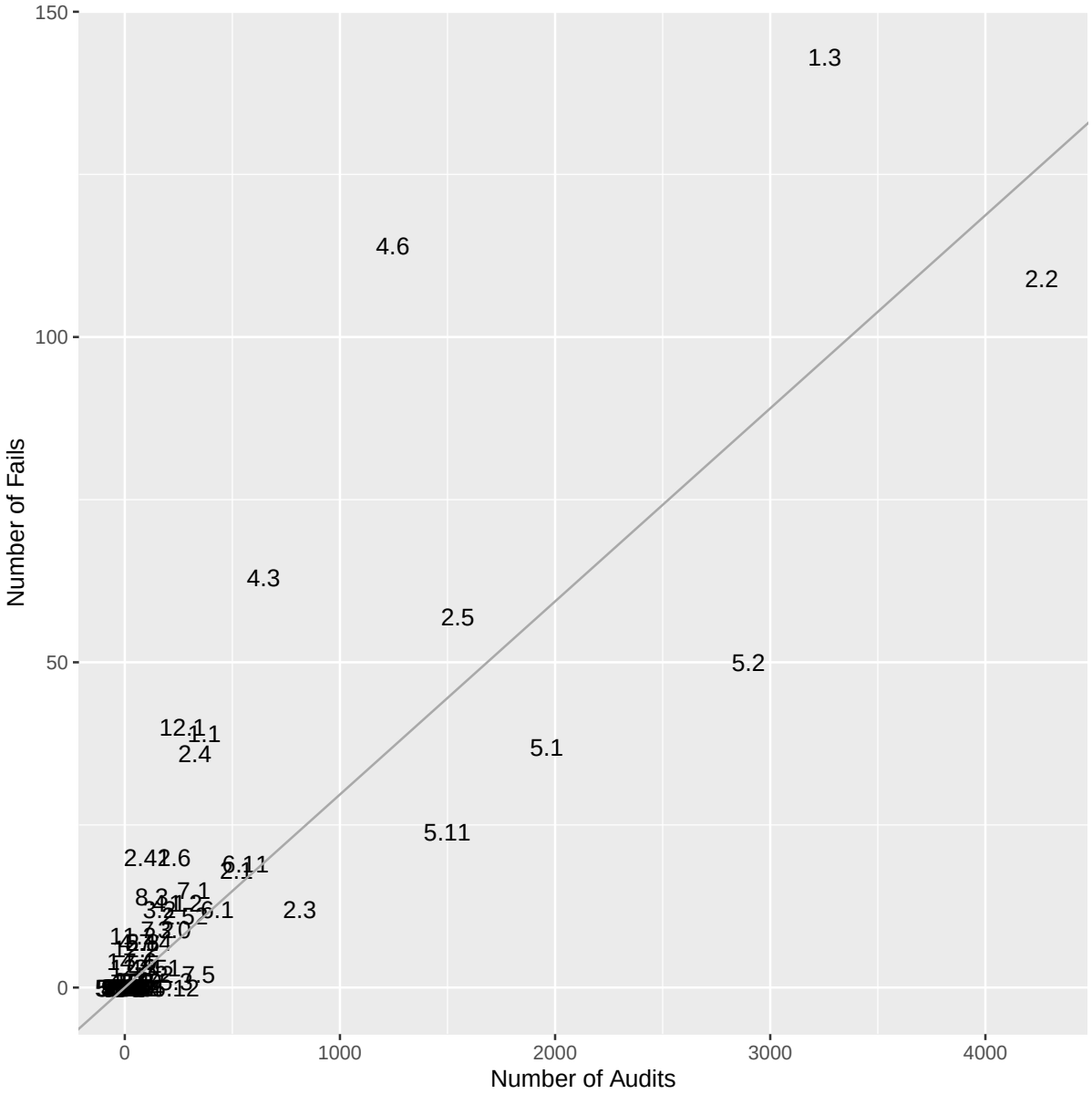

Figure B.1.: Class-specific count of audits and failed audits. Each class is represented by its identifier.

B.3. Methods

In order to allow the greatest possible flexibility in modeling the change in the failure probability as a function of the time since last audit, we applied generalised additive models to the data (Wood, 2017). These are models that allow the relationship to be represented by a smooth line that allows for some wiggliness but applies a penalty to try to induce smooth behaviour.

B.4. Results

An example of the kind of model that arises is presented in Figure B.2. We see peculiar phenomena such as a peak in failure rate at very low durations, which is hard

to understand and does not appear in the primary analysis (see Appendix A). Some patterns are common among the classes but these do not seem to support reasonable process interpretations, for example, the trajectories show unexpected wiggliness and many have a spike at zero that does not appear to be well supported in the data.

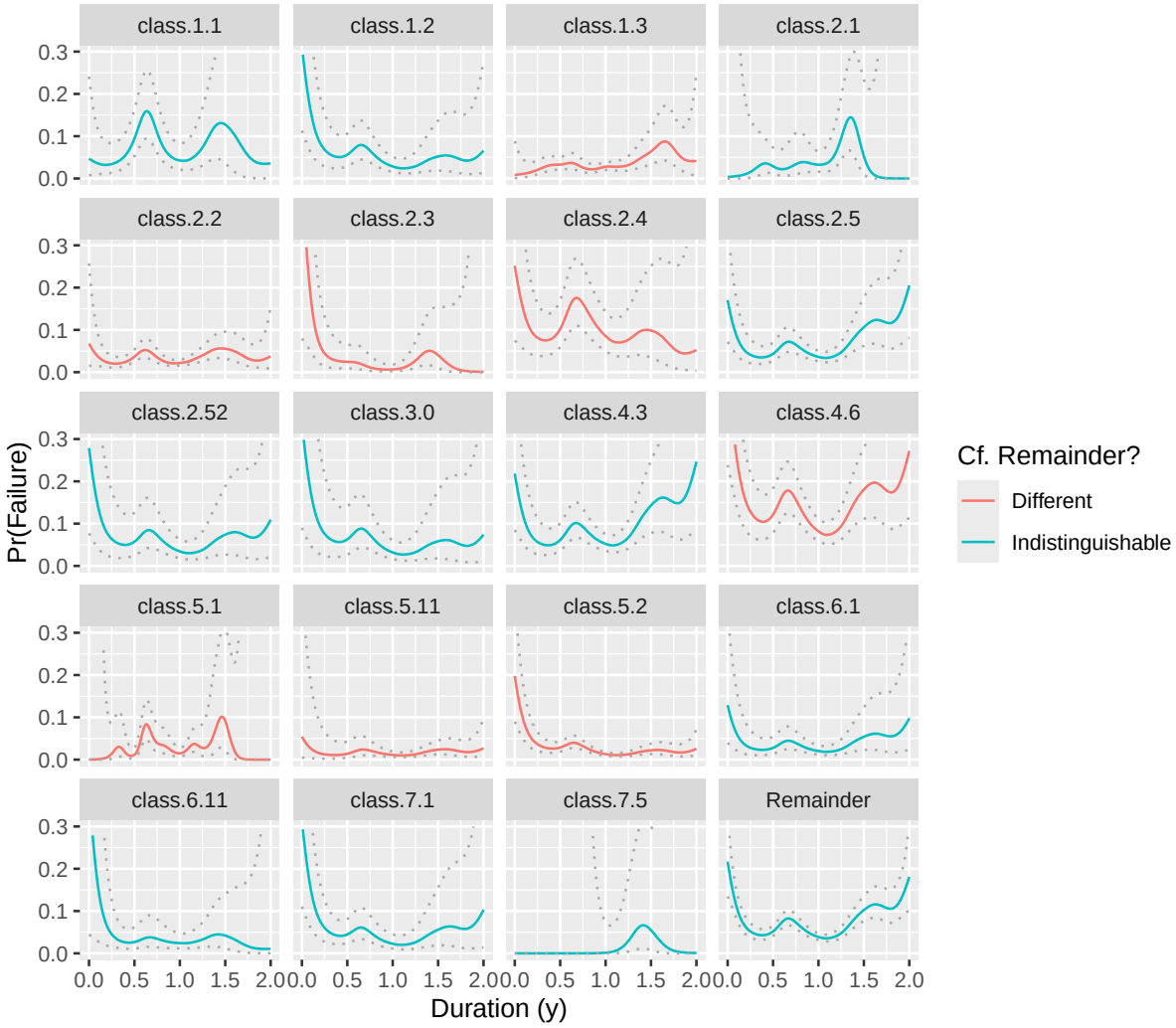

Figure B.2.: Modelled proportion of class-level audit failures, by class, as a function of time since last audit. The blue lines represent classes that are statistically indistinguishable from the aggregate of all the classes not represented here ('Remainder'). The red lines represent classes that are statistically different from the other classes. The dotted lines represent approximate 95% confidence intervals around the fitted lines.

B.5. Conclusions

We recognised that this line of analysis was becoming unfruitful, so we decided not to continue to try to develop class-specific models. Further work, which might involve different modeling assumptions or specifications, is certainly possible. However, in any case it is unlikely that the department would deploy a predictive class-based model. If class-based profiling were warranted then it is more likely that the

department would deploy an AA-specific compliance-based intervention approach,[1], and initiate the process using a class-independent algorithm such as developed in Appendix A. Such a scheme could cover variability based on class but also other factors that may be harder to measure.

[1]See, for example, the department's Compliance-Based Intervention Scheme, CBIS (https://www.agriculture.gov.au/biosecurity-trade/import/goods/plant-products/risk-return)

C. Criterion-specific analyses

C.1. Background

Here we report another analysis approach that we considered to be a dead end: criterion-level outcomes. As noted in the report, it is plausible to consider the criteria to be the fundamental observations. The criteria would then be clustered within audit, which would in turn be clustered within AA. An advantage of this approach is that it wouldn't matter which classes were being audited against, as each criterion would have its own failure probability and the difficulty of achieving compliance for the class could then be estimated for each class, based on their criteria. A summary of the number of times each criterion has been assessed against the rate at which the criterion has failed is presented in Figure C.1. Each point is the criterion reference number, which can be accessed by zooming the PDF (or, in the case of hard-copy, by means of a carefully deployed magnifying glass).

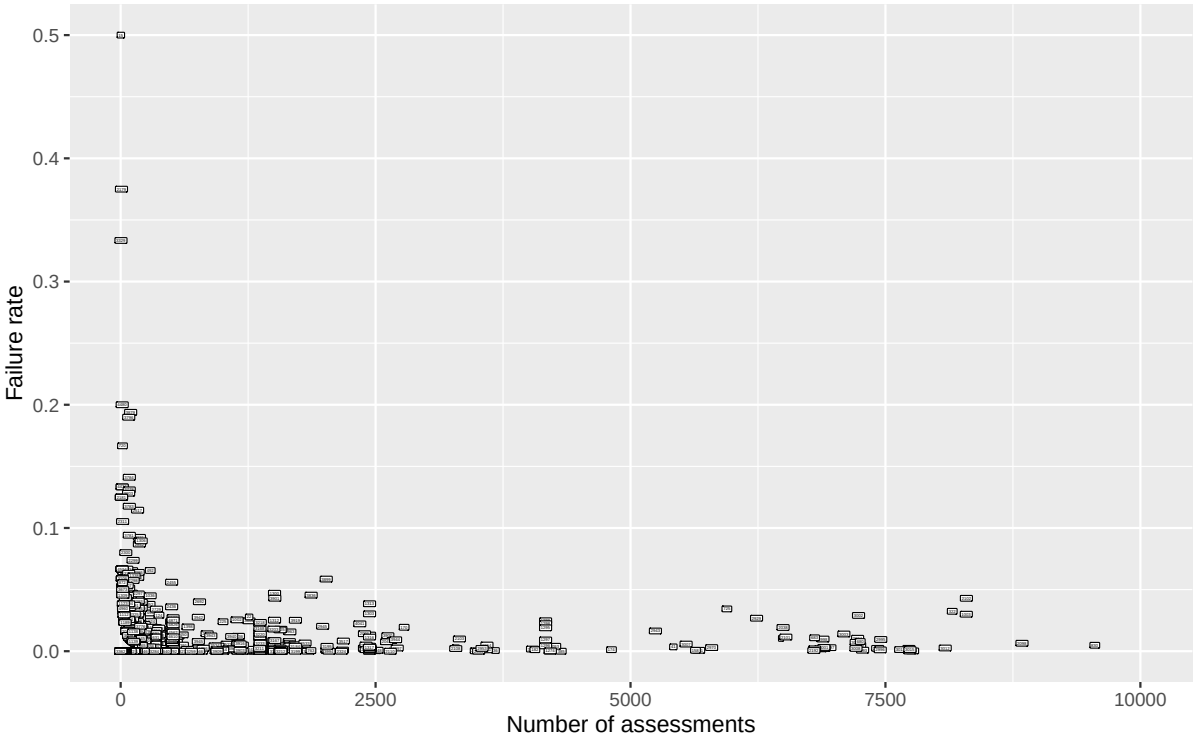

Figure C.1.: Proportion of failures for each criterion against the number of times the criterion has been assessed.

A wrinkle is that the database that records the criteria against the class[1] includes historical criteria as well, and we are unaware of any way to identify when the criteria

[1] namely, QPR.refClassConditions

dropped out of the class requirements. Consequently, we only know the complete set of criteria that ever applied to a class, which is extensive in some cases (Figure C.2).

C.2. Modeling

The modeling approach developed a *generalised linear mixed-effects model* (GLMM, Bates *et al.*, 2015; Stroup *et al.*, 2024) to predict the failure of a criterion. The binary response variable was whether the criterion assessment returned a fail. Final model terms comprised a term for whether the audit was announced or unannounced, a flexible function[2] of time elapsed since last audit in years, a baseline fail rate estimated at duration 0 (i.e., immediate re-assessment, expressed as an intercept in the model), and terms to represent the variation between different criteria (which is of primary interest), audits (in order to allow for within-audit failure patterns), and approved arrangements (in order to allow for across-audit within-arrangement patterns). The model also allowed for the time elapsed to have a different impact depending on the criterion and the audit, to allow for within-audit correlation.

Only the criteria corresponding to the second and any subsequent audits of approved arrangements were included, in order to be able to calculate the time elapsed since the previous audit.

C.3. Results

The predicted failure probabilities as a function of time since last audit are represented in Figure C.3. The panels of the figure each represent a class of AA, and the lines each represent a criterion within the class, that is, one of the criteria that is checked as part of the auditing of AA's that follow that class. Note that the criteria that apply to a class change from time to time and it is difficult to infer from the available data when criteria were relevant for a class.

C.4. Post-processing

The model makes predictions of criterion failures. In order to learn about the probability that audits of certain classes record a fail, we needed to aggregate the criterion-level results to represent the classes.

In order to do so we have to assume that the assessment of a criterion is unrelated to the class within which it is being assessed, that is, a failure for a criterion is established at the same level of non-compliance for all classes that require the criterion.

Naively we might expect that we could, for a given class (and therefore collection of criteria) and time elapsed, use the model to predict the criterion failure probabilities and suitably aggregate the failure probabilities up to the class level, much as one computes the probability of getting at least one six in a succession of four die tosses as $1-(1-\frac{1}{6})\times(1-\frac{1}{6})\times(1-\frac{1}{6})\times(1-\frac{1}{6}) = 1-(\frac{5}{6})^4 = 0.52$. However, the current setup is complicated for two reasons.

[2]Namely, a cubic spline

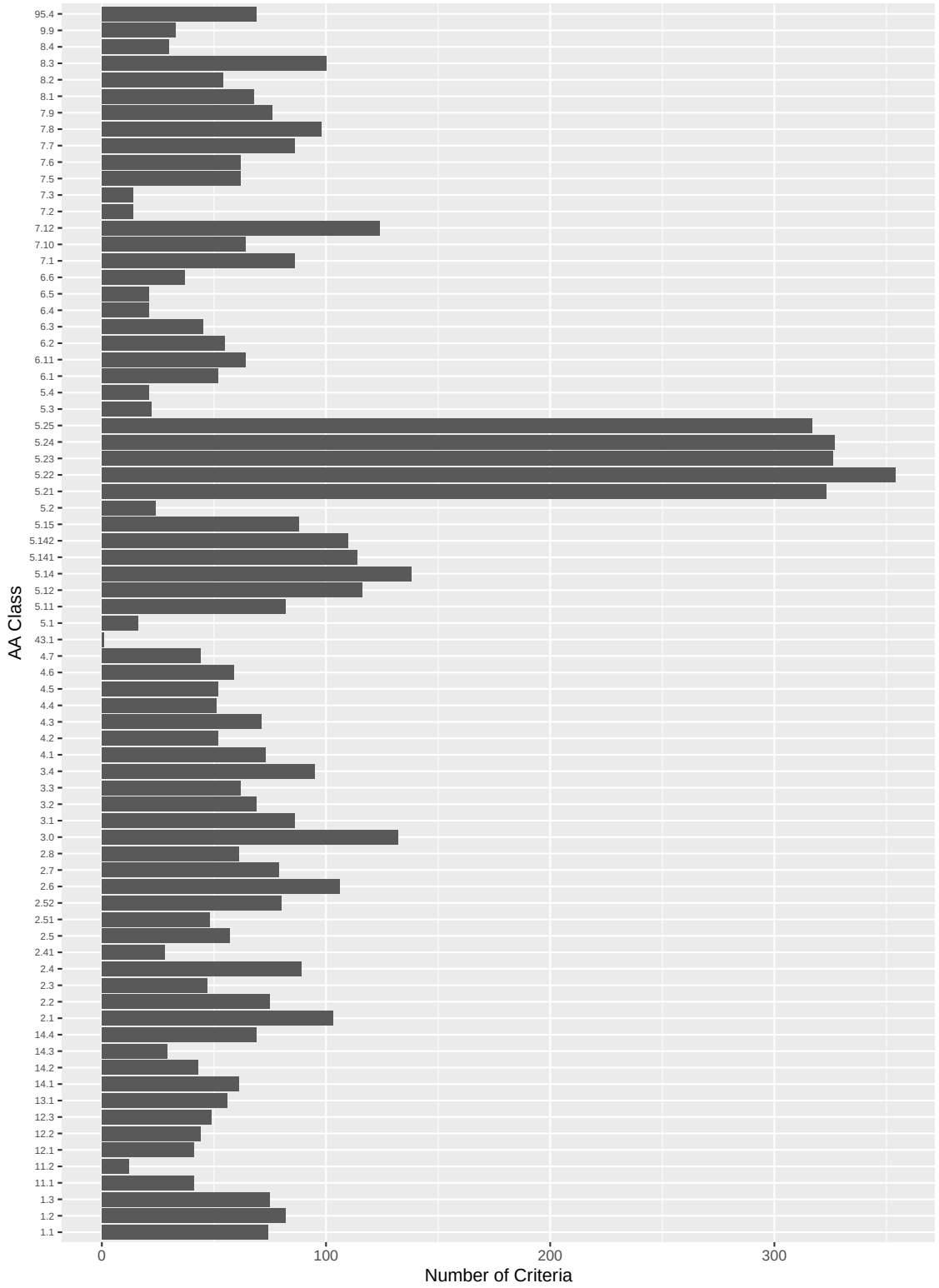

Figure C.2.: Total number of criteria that have ever applied to each AA class. The database includes historical criteria, for example criteria that have been edited and hence split into two: the original and the new criterion, the latter of which is assigned a new label, as well as criteria that are no longer applied.

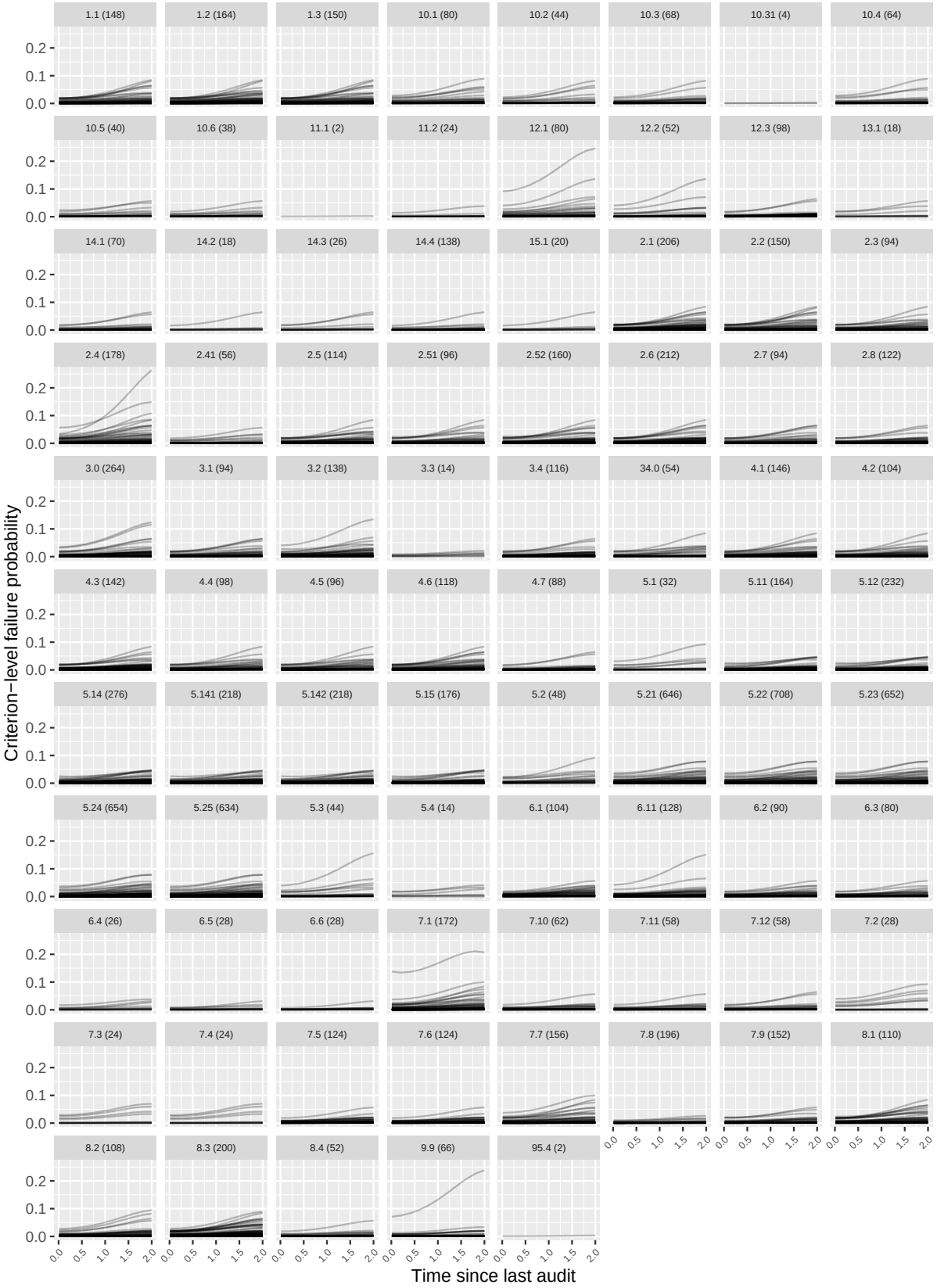

Figure C.3.: Modelled proportion of failures arising from unannounced audits for each criterion by time, clustered by AA class. Each line represents a criterion. Many criteria pertain to more than one class.

First, there is a the possibility that, within an audit, there may be intra-class correlation, *i.e.*, given at least one criterion fails within the audit, the others are more likely to fail.

Second, because the model applies shrinkage to the criterion-level trajectories, it will always predict a non-zero probability of failure at the criterion level, and many of the classes comprise many criteria, a problem that is amplified by our not knowing when criteria dropped out of the class definition. This is related to the so-called Birthday Paradox, in which it is startlingly probable that at least one of a sufficiently large number of outcomes will occur even if each occurs with seemingly negligible probability. Figure C.2 shows us that there are many criteria in most of the classes (a number of which will no longer be relevant) and when we aggregate even small probabilities across large numbers of experiments, the overall probability of an outcome builds up quickly. We computed the class-level failure probability for thresholds of 1 and 3 criteria for announced and unannounced audits; these are represented in Figure C.4. These results provide unrealistically high class-level failure probabilities, meaning that the failure rates of audits for particular classes greatly exceed the observed failure rates at the audit level (see, e.g., Figure A.1).

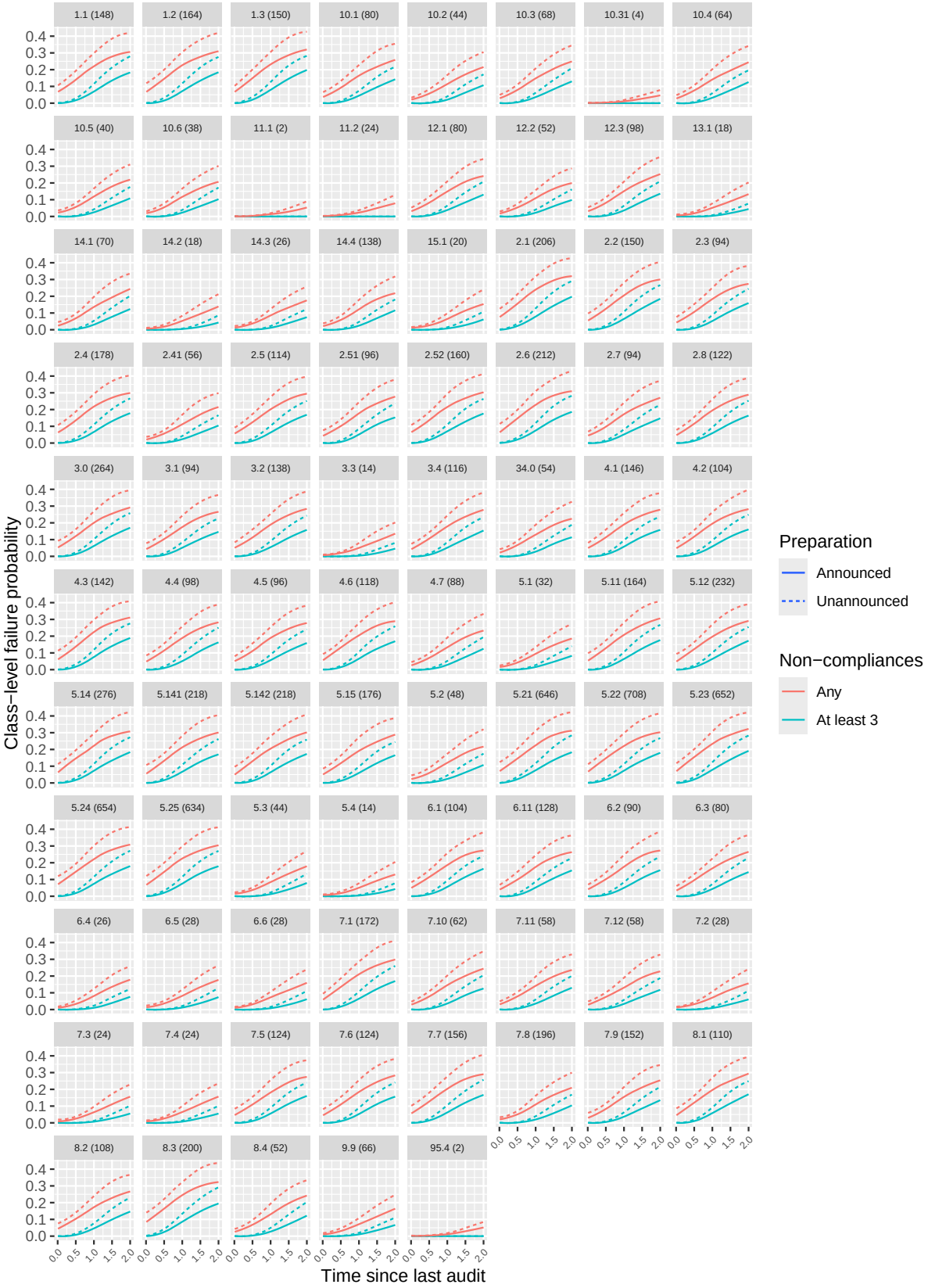

Figure C.4.: Modelled proportion of failures for each class by time for announced and unannounced audits. Many classes share common criteria. The line colour shows the failure criterion (one or three major non-compliances) and the line type shows the audit preparation (announced is solid; unannounced is dashed).