UNIVERSITY OF TARTU
School of Economics and Business Administration
Innovation and Technology Management Curriculum

Defir Ilmi Faridha

From Data Silos to Integrated Assurance: A Synthetic Data Approach to the Three Lines Model for Governance, Risk, and Compliance

Master's Thesis (18 ECTS)

Supervisors:
Mahmoud Shoush, PhD
Eduardo Ribas Brito, MSc

Tartu 2026

From Data Silos to Integrated Assurance: A Synthetic Data Approach to the Three Lines Model for Governance, Risk, and Compliance

Abstract:

The 2020 update of the Three Lines Model (3LM) aimed to shift Governance, Risk, and Compliance (GRC) from siloed control toward a more collaborative model. However, in practice, this transition is constrained by privacy regulations such as General Data Protection Regulation (GDPR), which limit cross-functional access to operational data. For example, when a compliance officer suspects money laundering, they typically cannot directly access the necessary transaction-level records. Instead, they must submit a formal request to IT and wait—sometimes for days—while sensitive customer data is extracted and handled across multiple steps. This creates what we refer to as the *governance–privacy paradox*: effective compliance requires data access, privacy rules restrict it, and organizational silos remain largely unchanged.

This study investigates whether synthetic data—artificially generated data that preserves statistical patterns without containing real personal information—can resolve this tension by enabling privacy-preserving yet analytically useful data sharing across governance functions. We simulate a banking GRC environment using two architectures: a traditional segmented Three Lines of Defense (3LoD) model, where data access is IT-mediated and siloed, and a collaborative (3LM) model, where synthetic data is shared instantly across all governance lines. Synthetic customer and transaction data are generated using a diffusion model and validated for statistical fidelity, analytical utility, and privacy preservation.

The results show that the collaborative model eliminates all routine IT-mediated data requests, removes exposure to personally identifiable information, and reduces investigation delay from 0.6 days to real-time access, while maintaining compliance and escalation integrity under deterministic simulation conditions. Instead of relying on IT as a bottleneck, all governance functions—risk, compliance, audit, and business—can access and analyze the same synthetic dataset concurrently. These improvements are consistent across three dimensions: information

granularity (data detail and accessibility), operational escalation (manual workload), and structural dependency (inter-functional reliance).

Keywords: synthetic data generation; Three Lines Model (3LM); governance, risk and compliance (GRC); privacy-preserving data sharing;

CERCS: P170, P175, S190

Andmehoidlatest integreeritud kindlustandva tegevuseni: sünteetiliste andmete põhine lähenemine kolme liini mudelile valitsemise, riskijuhtimise ja vastavuskontrolli valdkonnas

Lühikokkuvõte:

Kolme liini mudeli (3LM) 2020. aasta uuenduse eesmärk oli nihutada juhtimise, riski ja vastavuse (GRC) valdkond eraldatud kontrollilt koostööpõhisemale mudelile. Praktikas piiravad seda üleminekut aga privaatsuseeskirjad, näiteks isikuandmete kaitse üldmäärus (GDPR), mis piiravad valdkondadevahelist juurdepääsu operatiivandmetele. Näiteks kui vastavuskontrolli ametnik kahtlustab rahapesu, ei pääse ta tavaliselt otse juurde vajalikele tehingutaseme andmetele. Selle asemel peab ta esitama IT-osakonnale ametliku taotluse ja ootama mõnikord päevi, kuni tundlikke kliendiandmeid ekstraheeritakse ja töödeldakse mitmes etapis. See loob nn juhtimise ja privaatsuse paradoksi: tõhus vastavuskontroll nõuab juurdepääsu andmetele, privaatsuseeskirjad piiravad seda ning organisatsiooniliselt eraldatud andmed jäävad suures osas samaks.

See uuring uurib, kas sünteetilised andmed – ehk kunstlikult loodud andmed, mis säilitavad statistilisi mustreid ilma tegelikke isikuandmeid sisaldamatta saavad selle pinge lahendada, võimaldades privaatsust säilitavat, kuid analüütiliselt kasulikku andmete jagamist juhtimisfunktsioonide vahel. Simuleerime panganduse GRC–keskkonda, kasutades kahte arhitektuuri: traditsioonilist segmenteeritud kolme kaitseliini (3LoD) mudelit, kus andmetele juurdepääs on IT-vahendatud ja eraldatud, ning koostöömudelit (3LM), kus sünteetilisi andmeid jagatakse koheselt kõigi juhtimisliinide vahel. Sünteetilised kliendi- ja tehinguandmed genereeritakse difusioonmudeli abil ning valideeritakse statistilise täpsuse, analüütilise kasulikkuse ja privaatsuse säilitamise osas.

Tulemused näitavad, et koostöömudel kaotab kõik tavapärased IT-vahendatud andmepäringud (100% vähenemine), välistab juurdepääsu isikut tuvastavale teabele ja vähendab uurimise viivitust 0,6 päevalt reaalajas juurdepääsule, säilitades samal ajal 100% vastavuse ja eskalatsiooni terviklikkuse. IT-le kui pudelikaelale tuginemise asemel saavad kõik juhtimisfunktsioonid – riski-, vastavus-, auditi- ja ärijuhtimine – pääseda samaaegselt juurde samale sünteetilisele andmekogumile ja seda analüüsida. Need täiustused on järjepidevad kolmes dimensioonis: teabe detailsus (andmete detailsus ja juurdepääsetavus), operatiivne eskalatsioon (manuaalne töökoormus) ja struktuurne sõltuvus (funktsioonidevaheline sõltuvus).

Need tulemused näitavad, et sünteetilised andmed saavad lahendada juhtimise ja privaatsuse paradoksi, lahutades analüütilise juurdepääsu tundlike reaalsete andmete kättesaadavusest. Uuring näitab, kuidas kolme liini mudeli koostööl põhinevat eesmärki saab praktikas rakendada, pakkudes reguleeritud tööstusharudele skaleeritavat teed integreeritud tagatise saavutamiseks ilma privaatsusnõudeid kahjustamata.

Võtmesõnad: sünteetiliste andmete genereerimine; kolme liini mudel; valitsemine, riskijuhtimine ja vastavuskontroll; privaatsust säästev andmete jagamine

CERCS: P170; P175; S190

Contents

List of Figures

List of Tables

1. Introduction

1.1 Problem Context and Motivation

The Institute of Internal Auditors (IIA) introduced the updated 3LM in 2020, marking a shift from the traditional, siloed 3LoD approach toward a more integrated and collaborative GRC structure [1]. The updated model emphasizes coordination across IT, business operations, risk management, compliance, and internal audit functions to improve organizational assurance, transparency, and value creation.

Despite this shift in governance thinking, most organizations still operate under traditional 3LoD structures [2]. A key reason for this is the tension between integrated assurance and data protection requirements [3]. For many organizations, maintaining traditional structures— due to resistance to organizational transformation — is seen as a "safe" strategy to minimize operational costs and regulatory exposure, even though it creates inefficiencies that hinder organizational alignment [1, 4]. While the 3LM relies on shared access to operational data across functions, regulatory frameworks such as the General Data Protection Regulation (GDPR), the Digital Operational Resilience Act (DORA) [5], and Anti-Money Laundering (AML) directives impose strict limitations on the access and sharing of sensitive data [6, 7].

This creates a tension between data utility and compliance. Operational data, such as customer transactions, behavioral patterns, and compliance logs, is required by all three lines to perform their roles effectively [2]. However, in practice, access to such data is fragmented. The first line (IT, Data operations, and business operations) holds detailed transactional data, the second line (risk management and compliance) typically receives aggregated or filtered reports, and the third line (internal audit) relies on sampled or IT-mediated extracts [2, 3, 7]. Each request for detailed data introduces delays, operational overhead, and potential privacy risks [4].

For example, when the business operation identifies a potentially suspicious transaction pattern, accessing the underlying data often requires formal IT requests and approval workflows to be recursively invoked. This process can take hours or days and may expose large volumes of personal data, even when only a small subset is needed for investigation. As a result, risk signals may be analyzed too late to support timely intervention. The limitation, therefore, is not primarily technical but structural, arising from the underlying data access architecture.

The core problem addressed in this thesis is therefore: *How can organizations enable real-time, collaborative governance, risk, and compliance assurance while preserving data privacy and complying with regulatory constraints?*

1.2 Research Gap

Research relevant to this problem has developed along two largely separate streams: GRC on one side, and SDG on the other.

In the GRC and internal audit literature, studies consistently highlight persistent challenges, including fragmented risk information, siloed organizational structures, inconsistent data quality, and limited liquidity of information across lines [8, 9]. Despite the introduction of modern frameworks such as 3LM, organizations still rely on manual, periodic, and reactive assurance processes [1, 2, 4, 10]. Traditional risk assessments are typically conducted quarterly or annually, with ad hoc investigations initiated only after an incident occurs. This limits the ability to detect emerging risks or anomalies in a timely manner [1, 2, 10, 11].

Regulatory Technology (RegTech) solutions have improved specific operational tasks such as Know Your Customer (KYC), AML alert handling, and compliance documentation [8, 9]. However, they do not address the deeper structural issue of organizational alignment, and governance-wide data access [1, 4, 10]. As a result, each function within the 3LM typically operates with separate systems and datasets, leading to asymmetric information, redundant analyses, inconsistent interpretations, and inefficient assurance processes.

A traditional approach is no longer adequate for addressing the multifaceted nature of modern risks. Most organizations respond to these challenges by establishing separate functions to address specific risks [4]——exacerbates structural complexity and fragmentation rather than resolving the underlying root causes. Consequently, a paradigm shift is required; providing integrated assurance helps overcome the rapid evolution of risks driven by technological trends. In this regard, leveraging holistic, unified assurance is a pivotal strategy for organizations to strengthen governance and improve resilience and adaptability in an increasingly volatile business environment [4].

In parallel, SDG research has demonstrated strong potential to enable privacy-preserving analytics while maintaining statistical utility for machine learning tasks such as fraud and anomaly detection [7, 12]. However, these approaches are primarily designed for predictive modeling

and data analysis tasks rather than governance, audit workflows, or organizational assurance processes.

Importantly, these two research streams remain disconnected. GRC literature does not provide governance-wide technical mechanisms for secure, privacy-preserving information liquidity data sharing, while SDG literature rarely considers governance structures or assurance workflows as core design requirements. Even regulatory discussions, such as those from the Financial Conduct Authority (FCA), primarily focus on synthetic data for model validation and privacy-enhancing technologies rather than its role in enabling an integrated governance and assurance framework [7].

As a result, there is limited understanding of how synthetic data can be used beyond model training or analytical tasks, particularly as an enabler of integrated assurance across 3LM. This gap motivates the need to investigate whether synthetic data can serve as a governance infrastructure layer, enabling controlled, privacy-preserving access to operational data across GRC functions.

This study addresses this gap by positioning SDG as a mechanism of GRC integration rather than a purely technical tool.

More broadly, this study proposes that governance and assurance frameworks should be understood not only as organizational structures but as **information-access architectures** —— because without seamless information flow, governance becomes performative rather than functional. Adopting this perspective, this thesis evaluates how synthetic data can enable all functions to access analytically usable data despite privacy constraints, without compromising governance effectiveness. Ultimately, the focus centers on the downstream assurance impact of these information flows.

1.3 Research Questions

Based on the identified gap, this thesis is guided by the following main research question: *Can synthetic data resolve the governance–privacy paradox by enabling secure, collaborative, and analytically useful data access across all three lines of the Three Lines Model?*

With three sub-questions designed to deepen the scope of this investigation:

SQ1: Can synthetic data, as a governance enabler, eliminate silos and reduce reliance on the handoff data process?

SQ2: Does the transition from traditional to a synthetic data enabler improve data quality while maintaining or enhancing the completeness and consistency of risk-related information and analysis across functions?

SQ3: Can decentralized data access, by implementing a synthetic data model for 'self-serve', fully preserve compliance with SOPs and case escalation integrity while expanding data visibility across all governance layers?

1.4 Contributions

This study makes four distinct contributions – conceptual, methodological, empirical, and practical – each addressing a specific aspect of the research question.

1. **Conceptual contribution.** The study introduces synthetic data as a governance infrastructure layer for the 3LM. This research positions it as a structural mechanism that enables controlled, privacy-preserving cross-functional data access while maintaining organizational boundaries and regulatory compliance. By providing a shared source of consistent information across business, risk, compliance, and audit functions, synthetic data helps harmonize regulatory frameworks (e.g., GDPR, DORA, AML) with the collaborative intent of the 3LM. This reframing of synthetic data as a structural enabler, rather than a purely technical tool, is the core conceptual advance.

2. **Methodological contribution.** The study develops a simulation-based comparative strategy for evaluating governance processes under traditional (3LoD) and synthetic-data-enabled (3LM) architectures. This approach allows systematic, controlled comparison of different GRC configurations on identical operational data under deterministic rules. It operationalizes a set of governance-oriented metrics – information granularity, operational escalation, and structural dependency – that focus on architectural behavior rather than predictive accuracy. The methodology provides a reusable framework for assessing how changes in data access architecture affect downstream assurance activities, thereby offering an entry point for intelligent assurance in digital transformation contexts.

3. **Empirical contribution.** The study provides the simulation-based evidence comparing traditional 3LoD processes with a synthetic-data-enabled governance model. The

comparison quantifies improvements across three dimensions: data accessibility (from single-line IT mediation to concurrent access for all four functions), data consistency (from fragmented views to a single "golden source"), and cross-functional collaboration (from sequential handoffs to parallel self-serve investigation). The results empirically demonstrate that synthetic data can eliminate structural bottlenecks while preserving governance quality.

4. **Practical contribution.** The study offers actionable guidance for organizations seeking to implement the 3LM without violating privacy regulations. It demonstrates how synthetic data supports role-based access to analytically useful datasets, enabling real□time, self-serve analytical access across governance lines while keeping raw data restricted to evidence-level requests. The pipeline, validation metrics, and policy separation between routine synthetic access and exception-based raw access provide a scalable pathway for regulated industries to achieve integrated assurance.

More broadly, this study proposes that governance frameworks should be understood not only as organizational structures but as information-access architectures — because without seamless information flow, governance becomes performative rather than functional.

1.5 Scope and Limitations

This research is limited to governance, risk, and compliance processes within organizations structured according to the Three Lines Model. The focus is specifically on structured operational transaction data and its use in analytical and assurance workflows.

The proposed approach relies on synthetic data generation and does not consider alternative privacy-enhancing technologies such as homomorphic encryption or secure multi-party computation. The evaluation is conducted using a simulation-based design rather than a real-world enterprise deployment; this is a deliberate methodological choice suited to controlled comparison, though it means that the findings reflect experimental conditions rather than production-scale environments.

This study relies on raw operational data as the primary source to generate a representative synthetic dataset. Instead of pre-labeling the source data, deterministic mathematical logic for identifying suspicious transactions is applied directly during the simulation phase to ensure consistent risk detection across both AS-IS and TO-BE models.

It focuses on functional aspects such as data access, collaboration, and governance effectiveness, rather than legal interpretation or enforcement mechanisms.

1.6 Thesis Structure

This study is organized into six chapters. Chapter 2 reviews the background and related work, covering GRC and integrated assurance, the evolution from 3LoD to 3LM, suspicious transaction monitoring, and synthetic data generation. Chapter 3 presents the research design, including governance modeling, data access representation, simulation scenario, behavioral transformation, conceptual performance dimension, and the SDG pipeline. Chapter 4 reports the experimental setup, dataset, and sampling outcomes, evaluation metrics, baseline, evaluation logic, data quality-synthetic data validation, AS-IS and TO-BE simulations, governance simulation result metrics, comparative governance analysis, and summary. Chapter 5 discusses the findings, governance transformation, and the role of synthetic data, governance implications, organizational readiness, the relation to prior work, and practical recommendations. Finally, Chapter 6 summarizes the limitations and future work.

1.7 Use of AI Writing and Coding Assistants

Artificial Intelligence (AI)-assistants were used during thesis writing and code development as text-processing and formatting tools, in accordance with Section 3.2.2 of the Institute of Computer Science guidelines. Table 14 explains where and how these tools were used.

2. Background and Related Work

2.1 Governance, Risk, and Compliance (GRC) and Integrated Assurance

GRC frameworks emerged in response to major corporate scandals such as Enron and WorldCom in the early 2000s, which exposed weaknesses in organizational governance, accountability, and risk oversight [13]. The concept was formally introduced by the Open Compliance and Ethics Group (OCEG) as a structured approach for integrating governance, risk management, and compliance activities across organizations [14]. Since then, GRC has evolved into a widely adopted best-practice framework for improving organizational sustainability, operational resilience, and regulatory compliance [14, 15].

Traditionally, GRC initiatives focused primarily on regulatory compliance and enforcement of controls [1, 14, 16]. However, modern GRC frameworks increasingly position governance and assurance as strategic capabilities that support organizational transparency, accountability, operational efficiency, and stakeholder trust [1, 15, 17]. Recent updates to the Open Compliance and Ethics Group (OCEG) – GRC Capability Model and the IIA Three Lines Model emphasize continuous monitoring, integrated assurance, and interoperable governance rather than isolated or periodic compliance activities [1, 15].

This shift is particularly important in highly regulated sectors such as financial services and fintech, where organizations operate under complex regulatory frameworks including GDPR [5], AML directives[18], DORA[5], and Network and Information Security Directive 2 (NIS2)[5]. Effective GRC implementation in such environments depends on coordinated information sharing, consistent risk assessment, continuous monitoring, and collaboration across operational, compliance, risk, and audit functions.

Despite these developments, previous studies consistently show that GRC processes in practice remain fragmented, manual, and reactive [3, 4]. Many organizations continue to perform governance and assurance activities periodically, often through quarterly or annual assessments, with ad-hoc investigations triggered only after incidents occur [11]. Such approaches reduce the ability to detect emerging risks, suspicious activities, or compliance violations in a timely manner [11, 19].

One of the primary challenges limiting integrated assurance is fragmented data access across organizational functions. Business operations, compliance teams, risk management units, and

internal audit departments frequently operate using separate systems, datasets, and reporting structures [2, 3]. As a result, downstream assurance functions often receive incomplete, delayed, or heavily aggregated information, limiting visibility into operational risks and reducing the effectiveness of governance activities.

For example, operational teams may have direct access to detailed transaction records and customer behavior data, while compliance and audit functions rely on filtered reports or IT-mediated extracts. In practice, this means that suspicious activity investigations, risk assessments, or audit procedures may be delayed because access to the required operational data requires formal approval workflows [2, 10]. Such fragmentation increases operational costs, creates duplicated assurance activities, and weakens organizational alignment.

In daily transaction processing, first-line operational teams monitor and reconcile transactions. Business teams request transaction histories and behavioral patterns to conduct analyses such as RFM segmentation for customer targeting. At the same time, compliance and risk teams require access to the same data for identifying anomalies and suspicious activities. However, due to strict internal data access policies, IT departments may restrict access and provide only partial extracts or aggregated summaries, preventing full-population testing or validation of risk signals and resulting in incomplete conclusions. Similarly, an internal audit may be unable to obtain timely evidence due to data stored across multiple disconnected systems. These real-world constraints illustrate how limited data access directly impedes collaboration, slows assurance activities, and weakens the organization's ability to detect issues early.

Although RegTech solutions have automated specific operational tasks such as KYC, AML alert handling, and compliance reporting, these systems typically address isolated functions rather than enterprise-wide governance integration [8, 9]. Consequently, many organizations continue to experience siloed processes, duplicated testing efforts, inconsistent reporting practices, and limited information sharing across governance functions [6, 7].

These structural challenges are closely interconnected. Differences in data governance practices across departments complicate regulatory harmonization, while restrictive data access policies reinforce organizational silos [2, 3]. This creates a recurring cycle in which limited data sharing weakens governance integration, while the lack of integration further complicates effective data governance [4, 10].

To address these limitations, organizations increasingly require mechanisms that support information liquidity across lines and data accessibility without violating privacy and regulatory requirements. Within this context, the SDG mechanism could enable privacy-preserving data sharing across governance functions [6, 7]. By replacing sensitive operational data with statistically representative synthetic data, organizations may improve collaboration, transparency, and integrated assurance while remaining aligned with data protection principles such as GDPR data minimization [12, 20].

Table 1. GRC structural challenges in the traditional setting

Aspect	**Current Situation**	**Underlying Gap**
Siloed processes	Functions operate independently with limited collaboration; siloed systems and manual handoffs	Lack of coordinated, asymmetric information flow
Data sharing & privacy	Limited data sharing across departments; strict privacy rules restrict operational data access	Inability to provide consistent information across GRC functions
Delayed downstream impact	Downstream GRC functions receive incomplete or delayed data	Lack of early insights hinders proactive risk identification
Automation & integration	GRC tools remain fragmented and lack unified pipelines	No centralised data foundation supporting end-to-end GRC activities

This project began with the our observations and practical experiences. An organization that provides a shared analytical data repository across multiple organizational lines—even one deemed relatively "secure" due to the organization's established level of access to raw operational data—can become an entry point for a large-scale data breach.

The consequences of such a breach extend beyond regulatory sanctions to the complete loss of customer data integrity, leading to reputational risk and a decline in public trust in the company [21].

This incident could serve as a starting point for a shift away from traditional approaches. When leveraging synthetic data as an analytical data repository, mitigating this risk is needed to enable governance oversight activities without exposing privacy-sensitive records. Access to raw data can thus be restricted to only those required as evidence, in cases where evidence-level transaction details are needed for in-depth investigations through targeted, formally justified data requests. However, daily business and operational activities can be covered by the synthetic data.

This data access design principle aligns with the GDPR 's data minimization requirements and with the Three Lines Model's emphasis on role-based information boundaries and role separation to manage risks, foster accountability, and drive value, moving away from rigid, siloed "defenses" toward a more collaborative, role-based governance structure.

2.2 Governance Models: From the Three Lines of Defense (3LOD) to Three Lines Model (3LM)

For decades, organizations have implemented governance structures based on the 3LoD model [1]. Under this model, responsibilities are distributed across three organizational layers: operational management as the first line, risk and compliance management as the second line, and internal audit as the third line [1, 16]. The model was designed to clarify accountability, strengthen risk oversight, and maintain independence between assurance functions.

In practice, however, the strict separation of responsibilities within the 3LoD framework often led to siloed organizational structures [2]. Each line typically operated independently, maintaining separate systems, reporting mechanisms, and information boundaries [2, 10]. Although this separation strengthened formal independence, it also limited collaboration and reduced the efficiency of governance activities.

These limitations became increasingly problematic in data-driven and highly regulated environments. Internal audit functions frequently lacked timely access to operational transaction data held by business functions, while compliance and risk teams often duplicated analyses

because information was not effectively shared across lines [3, 4]. As a result, governance activities became fragmented, slow, and operationally inefficient [4, 11].

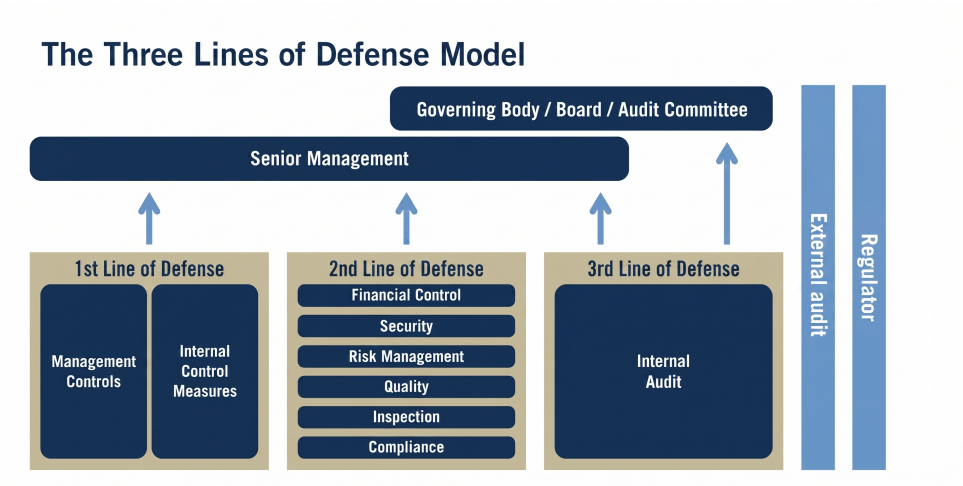

Figure 1. The Three Lines of Defense Model, illustrating the separation of responsibilities across three organizational layers: operational management (first line), risk management and compliance (second line), and internal audit (third line), with oversight from the governing body and external assurance providers [22].

To address these challenges, the IIA introduced the updated 3LM in 2020 [1]. Unlike the traditional 3LoD framework, the 3LM shifts the emphasis from *defense* toward collaboration, value creation, and integrated assurance [1, 10]. The updated model promotes communication, coordination, and shared accountability across governance functions while preserving the independence of internal audit [1, 2].

Under the 3LM, governance effectiveness depends on coordinated assurance activities supported by consistent and timely information flows across organizational functions. Still, this collaborative objective creates a practical governance paradox [2, 4]. Effective assurance integration requires interoperable access to operational data, including sensitive customer transactions, compliance logs, and risk assessment outputs [3, 10]. At the same time, regulations such as GDPR restrict the direct sharing of such information across organizational boundaries [5, 6].

Consequently, organizations face competing requirements. Governance frameworks increasingly demand integrated and collaborative assurance processes, while privacy regulations limit direct access to operational data [4, 5]. This tension creates notable challenges for implementing the collaborative intent of the Three Lines Model in practice [2, 3].

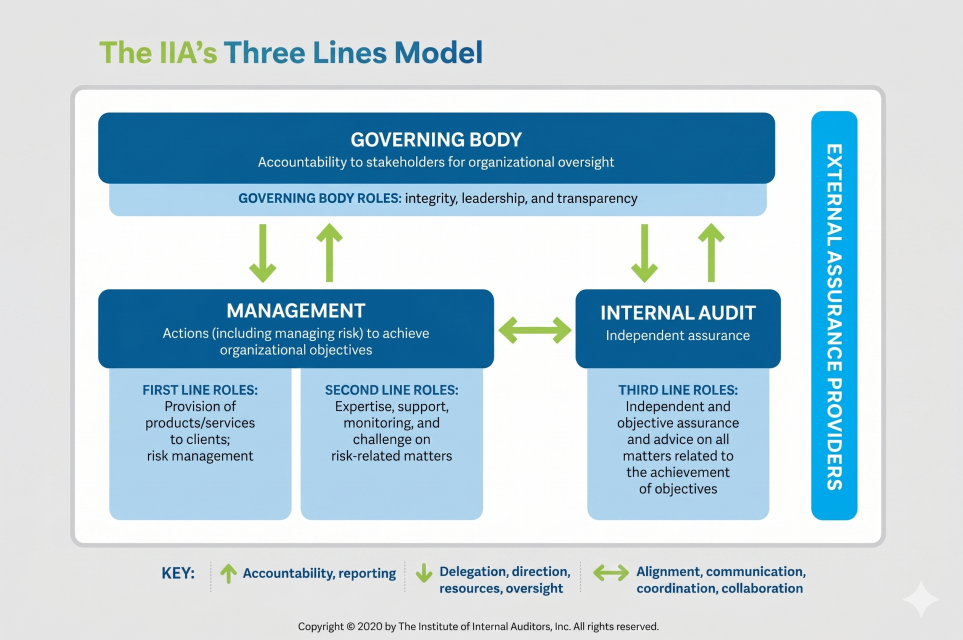

Figure 2. The IIA's Three Lines Model (2020), illustrating the shift from defensive separation toward collaborative integration across the governing body, management, and internal audit, with emphasis on communication, coordination, and shared accountability across all governance functions [1].

Within this context, Privacy-Enhancing Technology (PET) becomes increasingly important. Among these, SDG has emerged as a promising mechanism for enabling collaboration and integrated assurance without exposing Personally Identifiable Information (PII). By providing statistically representative datasets that preserve analytical usefulness while protecting privacy, synthetic data may help organizations reduce siloed processes, improve information accessibility, and support the collaborative objectives of the Three Lines Model.

This study adopts the 3LM as its primary governance lens to demonstrate how improved data sharing—enabled through SDG—can support the collaborative intent of the 3LM and mitigate common challenges such as silo mentality, restricted data access, and duplicated testing efforts [2, 10].

2.3 Suspicious Transactions as the Use Case

This thesis uses suspicious transaction monitoring as an illustrative operational use case to evaluate how improved data accessibility and cross-functional collaboration can support integrated governance and assurance. The focus of the study is not the development of

AML detection algorithms themselves, but rather the governance and data-sharing challenges surrounding suspicious transaction analysis.

From a governance perspective, suspicious transaction monitoring provides an appropriate case study because it inherently requires collaboration across multiple organizational functions [2, 4]. Operational teams monitor transactions, compliance teams investigate suspicious patterns, risk management assesses broader exposure, and internal audit evaluates the effectiveness of governance controls. However, these activities rely on access to sensitive customer data, making suspicious transaction monitoring particularly suitable for evaluating privacy-preserving data-sharing mechanisms such as synthetic data generation [6, 7].

Conventional AML systems typically define suspicious transactions using predefined rules and fixed thresholds (e.g., large cash deposits or specific cross-border transfers). This approach is typically binary and often yields a high false-positive rate [23].

In this study, suspiciousness is defined by the degree of deviation from a customer's historical behavioral profile rather than by absolute transaction values [23].

This deviation-based definition provides earlier signals of operational risk and aligns with the TO-BE objective of this study—shifting GRC processes from reactive, periodic assessments to more proactive, continuous monitoring[19, 23].

2.4 Synthetic Data Generation (SDG)

SDG has emerged as a prominent privacy-preserving technology for enabling data analysis without exposing sensitive information [6, 20]. Synthetic data refers to artificially generated data that reproduces the statistical properties, structure, and behavioral patterns of real datasets while avoiding direct disclosure of personally identifiable information [24, 25].

Over the past decade, SDG has gained significant attention across domains such as healthcare, finance, and cybersecurity, where strict privacy regulations limit access to operational data [12]. The increasing adoption of machine learning and AI systems has further accelerated interest in synthetic data as organizations seek ways to overcome data scarcity, support model development, and comply with privacy requirements [4, 7, 11, 26].

SDG typically involves learning the statistical characteristics of an original dataset (assuming we have access to it) and generating new artificial records that preserve the distributional properties

and relationships observed in the source data [24]. Various techniques have been proposed for this purpose, including Generative Adversarial Network (GAN)s, Variational Autoencoder (VAE)s, and diffusion-based models [27, 28].

Different sectors have adopted SDG technologies for different purposes. Healthcare applications commonly focus on patient records, rare disease research, and AI diagnostics, while financial applications emphasize fraud detection, credit scoring, and AML analytics. Regulatory and industry reports similarly focus on SDG as a privacy-enhancing technology for model development, testing, and validation [6, 7].

Despite these advances, existing SDG research remains highly fragmented across domains and use cases. Most studies focus on improving technical properties such as utility, privacy preservation, fairness, or model performance [20, 26]. Very limited attention has been given to integrated assurance, and organizational alignment [7, 12].

This fragmentation reflects a deeper disciplinary boundary: SDG development remains concentrated on technical aspects—such as increasing utility, reducing bias, and strengthening privacy — while governance aspects remain largely untouched. As a result, peer-reviewed literature addressing the cross-functional application of synthetic data within integrated governance frameworks remains scarce, particularly regarding its potential to enhance collaboration and assurance under the Three Lines Model. The absence of SDG–GRC research is a result of disciplinary separation between synthetic data research (computer science and statistics) and governance assurance research (accounting, internal audit, and regulation) than to the topic being irrelevant.

This study addresses this gap by positioning synthetic data not merely as a machine learning tool, but as a governance enabler that supports integrated assurance within the Three Lines Model. By applying SDG as a mechanism that bridges the trade-off between information access and data protection identified in Section 2.2, this research connects the technical capabilities of synthetic data with the organizational requirements for interoperable governance GRC integration.

2.5 Related Work

Prior research related to suspicious transaction monitoring and synthetic data generation can be grouped into several major streams.

First, behavioral and dynamic risk indicators for detecting suspicious activity. These studies demonstrate that changes in customer transaction behavior, such as shifts in frequency, transaction volume, or counterparties, provide stronger indicators of suspicious activity than static customer risk profiles [23]. Such approaches require continuous access to operational transaction data and, therefore, depend heavily on effective data access and real-time information sharing.

Second, real-time detection architectures using streaming technologies, unsupervised learning, and distributed processing systems. These studies show that streaming-based architectures significantly improve detection speed and responsiveness compared to traditional batch-processing approaches [19]. However, they primarily address technical optimization rather than organizational governance or assurance integration.

Third, graph-based transaction analysis. By modeling financial transactions as interconnected networks, these approaches enable the identification of complex laundering patterns that are difficult to detect using traditional flat transaction tables [29]. Although effective from a technical perspective, such systems remain largely confined to isolated operational or compliance functions rather than integrated governance environments.

Fourth, privacy-preserving synthetic data generation for AML and fraud detection systems. These studies typically apply GAN- or VAE-based methods to balance datasets, improve model performance, or enable privacy-preserving analytics [20, 27]. However, their objective remains improving machine learning performance within isolated analytical workflows rather than enabling integrated assurance.

Collectively, these research streams demonstrate considerable progress in transaction monitoring, anomaly detection, and privacy-preserving analytics [2, 3]. Reite et al. [23] identify the need for dynamic real-time data streaming, Cardoso et al. [30] highlight the importance of graph-based structures, and Chen et al. [29] demonstrate mechanisms for privacy-preserving data sharing through synthetic data generation. However, existing research remains primarily focused on technical performance and does not connect these capabilities to governance structures that determine who receives information, under what authority, and for what assurance purposes.

This gap between detection capabilities and governance integration is precisely the space this thesis addresses. By applying SDGs not only as a training mechanism but also as a governance

instrument within the three lines model, this research bridges the technical and organizational dimensions that the existing literature has largely treated separately.

3. Methodology

3.1 Research Design

This study adopts a simulation-based comparative research design to investigate how governance structures influence data accessibility, operational collaboration, and audit effectiveness in financial GRC processes.

Two governance configurations are analyzed:

1. **Segmented governance (AS-IS)** based on the traditional 3LoD model.
2. **Collaborative governance (TO-BE)** based on synthetic-data-enabled 3LM .

The purpose of this comparison is to examine how different data access designs—enabled by synthetic data—affect downstream operational behavior and information usability across the Three Lines Model. Rather than being focused on optimizing detection accuracy for suspicious transactions. The design illustrates the data flows and governance mechanisms that support early operational risk signals and enable grouping, review, and escalation within the GRC environment.

The research is structured as a controlled comparative experiment in which the governance model is the only manipulated variable, while the operational environment, transaction data, and behavioral assumptions remain consistent across both scenarios.

The methodology follows four conceptual phases: (1) modeling the segmented governance structure as a baseline using the original transaction dataset, (2) defining behavioral transformation rules for converting raw transactions into risk signals, (3) introducing a synthetic data-enabled collaborative environment as the TO-BE model, and (4) performing a comparative analysis of system-level outcomes across both governance configurations.

Each phase is described in detail in the subsections that follow.

The phase (1) focuses on establishing the baseline using the original transaction dataset to simulate a traditional GRC environment. In this state, data access requires IT mediation and a request-driven approach, reflecting a restricted, siloed environment in which non-IT/data operations functions rely on masked or aggregated summaries, leading to information asymmetry. The phase (2) introduces the behavioral logic for converting raw label transactions into risk signals (e.g., 1.5x and 2.0x deviation ratios). These rules are applied uniformly across both the original and

synthetic data environment to ensure a fair comparison and to support traceable grouping, review, and escalation. The phase (3) introduces synthetic data to enable a collaborative environment (TO-BE). In this phase, synthetic data is generated using diffusion model techniques. The aim of this state is to redesign the flow so that all lines have direct, autonomous access to this shared data environment, establishing synthetic data as an enabler that preserves the statistical structure of the original data while removing privacy constraints and bypassing traditional IT bottlenecks. The phase (4) is the final phase of the research design. performing a comparative analysis of system-level outcomes to evaluate the impact of the two governance layer designs. The evaluation focuses on system-level performance dimensions, specifically information usability, data accessibility, and the reduction of handoff friction, rather than algorithmic detection accuracy.

3.2 Governance Modelling Assumptions

The AS-IS model represents a segmented governance structure (3LoD) in which responsibilities are divided among operational, risk management, compliance, and audit functions. This separation creates controlled information silos, where each line has partial visibility into the data. Access to detailed information is mediated through formal requests and IT-controlled processes, creating hurdles to proactive risk identification.

The TO-BE model represents a collaborative governance structure (3LM), which replaces rigid separation with collaborative integration. In this model, all lines operate within a shared, controlled data environment that enables direct and structured access to consistent information while preserving privacy constraints. With this model, each line can establish integrated assurance and proactive action to create value.

The key assumption is that both models operate on identical underlying transactional behavior, and the only structural difference is the governance and data access architecture. This allows observed differences to be attributed to governance design rather than data variation.

3.3 Data Access Representation

While Section 3.2 defines the structural boundaries, this section specifies how **data access** is modeled as a structural property of governance rather than a technical feature of the dataset.

In the segmented setting, data access is request-driven and mediated by IT functions. Operational, risk management, compliance, and audit teams do not directly access raw transaction data. Instead, they rely on extracted or aggregated views of the data, which can lead to delays and

potential information loss. Furthermore, this condition can hinder taking action immediately to prevent operational risk.

In the collaborative setting, data access is enabled through a shared synthetic data layer. This allows all governance lines to access consistent, privacy-preserving representations of the same underlying behavior without requiring reconstruction through IT intermediaries, enabling continuous monitoring to support early risk detection.

The key distinction between the two models is therefore not data availability, but the degree of autonomy, latency, and dependency embedded in the access process.

3.4 Simulation Scenario Formalisation

The simulation is based on a single operational scenario: suspicious transaction behavior. This scenario is used as a proxy for general behavioral risk detection in financial systems.

Suspicious behavior is defined through deviations from customer-specific historical transaction patterns. Each customer is modeled using their own baseline behavior derived from historical transaction frequency and transaction value.

Deviation thresholds are defined as policy-based design parameters rather than regulatory constraints. Specifically, 1.5× and 2.0× deviation bands are used to categorize behavioral intensity. These thresholds represent increasing levels of deviation from expected behavior and are used consistently across both governance models to ensure comparability.

The purpose of this scenario is not to detect fraud, but to simulate how different governance structures interpret and escalate behavioral anomalies.

Parameter Selection Rationale

The threshold and baseline parameters applied in this study were determined through an iterative calibration process informed by both empirical observation and established AML monitoring guidance.

Deviation thresholds (1.5× and 2.0×). The initial threshold was considered at 1.2×, reflecting a 20% operational variance margin as an internal control design parameter based on risk appetite. However, a threshold of 1.2× was found to be overly conservative, resulting in a high volume of escalations that would generate excessive operational workload without a proportional risk-detection benefit. The threshold was subsequently set to 1.5× to provide a wider tolerance band

consistent with customer profile variability, in alignment with the risk-based approach to AML monitoring recommended by the Financial Action Task Force Financial Action Task Force (FATF) [18]. Transactions exceeding 2.0× the historical baseline are classified as significant deviations and trigger immediate escalation.

Minimum observation period (min_periods = 5). The expanding average baseline requires a minimum number of historical observations before deviation ratios can be computed. This parameter was calibrated using customers with irregular transaction behavior, as high-variability customers make the trade-off between baseline stability and detection responsiveness most visible. A minimum period of less than five days produced an overly reactive baseline with high false-positive rates, while periods above five days introduced detection delays that could allow anomalous behavior to persist undetected. Five days were selected as the balance point between baseline stability and detection responsiveness.

Exclusion of the observation day. The baseline is computed using only historical data strictly prior to the observation day ($\tau < t$). Including the observation day would introduce look-ahead bias — the current day's behavior would influence its own deviation score, artificially suppressing anomaly signals on days with genuinely suspicious activity.

3.5 Behavioural Transformation Model

Raw transaction data is transformed into structured behavioral signals through a multi-step abstraction process.

First, transactions are aggregated at the customer-day level to capture daily behavioral patterns. Second, customer-specific baselines are constructed using historical averages of transaction frequency and value. Third, deviation ratios are computed by comparing current behavior to historical baselines.

These deviation ratios are then mapped into categorical behavioral groups: normal, elevated, and significant. Finally, these groups are combined to generate overall suspicion labels: low, medium, and high, as illustrated in Figure 3.

Cases generated from observation data are then routed to Data Operations as part of its workflow design (Figure 4), which classifies each customer-day observation and routes it to the appropriate operational action. Cases flagged as high risk are then handled through the business operations escalation workflow (Figure 5), with the results documented in an audit

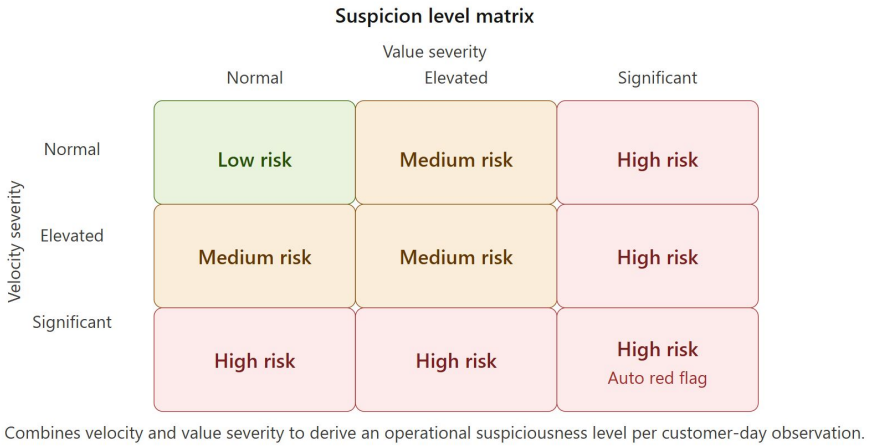

Figure 3. Suspicion level matrix — combining transaction velocity, severity, and value severity to derive an operational suspiciousness level per customer-day observation. Cells shaded in darker tones indicate higher risk classifications, with the Significant–Significant combination triggering an automatic red flag for immediate escalation.

trail. The auto-red-flag trigger is aligned to the same threshold as the significant severity level (ratio > 2.0), ensuring consistency across the monitoring pipeline. Any observation classified as significant simultaneously triggers an operational action, eliminating logical gaps between severity classification and escalation.

This transformation process ensures that raw transactional data is converted into interpretable behavioral signals that can be consistently analyzed across different governance structures. The objective is to enable traceable and comparable risk abstraction rather than predictive classification.

3.6 Conceptual Performance Dimensions

The evaluation framework is based on three conceptual dimensions that describe governance performance at the system level.

The first dimension is *information granularity*, which captures how much detail is preserved or lost as data moves across organizational structures. It reflects the level of abstraction introduced by governance design.

The second dimension is *operational escalation*, which captures the extent to which cases require manual intervention, cross-line coordination, or formal escalation procedures. It reflects operational workload and responsiveness.

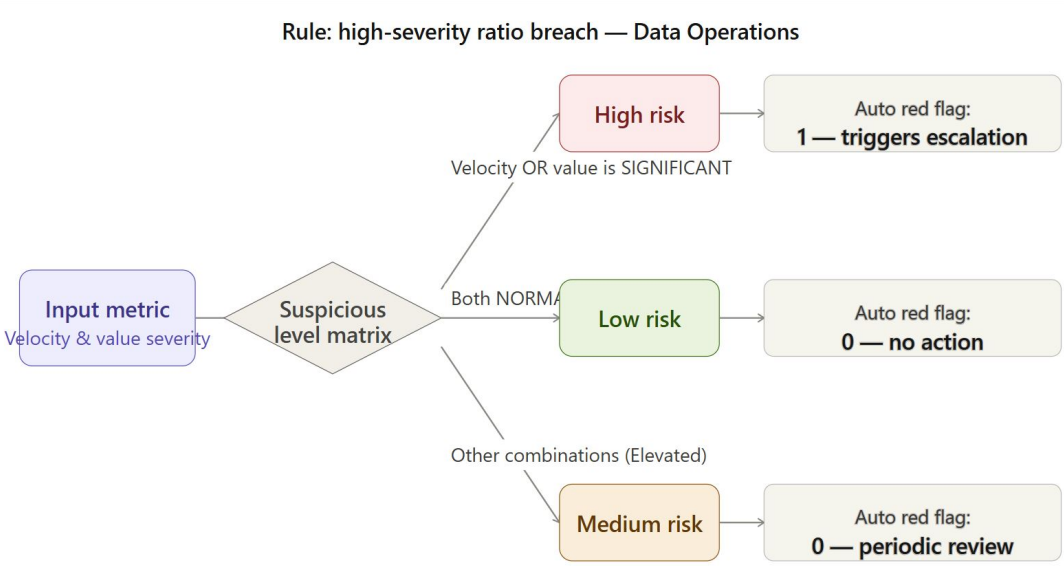

Figure 4. Data Operations pipeline for high-severity ratio breach detection. The pipeline classifies each customer-day observation using the suspicion level matrix and routes cases to the appropriate Operational action: escalation (high risk), periodic review (medium risk), or no action (low risk).

The third dimension is *structural dependency*, which captures the degree to which organizational functions depend on other units (particularly IT) to access or reconstruct information. It reflects systemic coupling and organizational friction.

These dimensions are used to evaluate governance efficiency rather than model accuracy or predictive performance.

3.7 Synthetic Data Generation as a Governance Abstraction

SDG is used as an enabling mechanism for the collaborative governance model rather than as a standalone modeling objective.

A diffusion-based tabular generative model i.e., Tabular Denoising Diffusion Probabilistic Model (TabDDPM), is used to produce synthetic datasets that preserve statistical structure while removing direct identifiers and sensitive dependencies. The role of synthetic data in this study is to enable controlled data sharing across governance lines without violating privacy constraints.

The synthetic data layer acts as an abstraction mechanism that decouples data utility from data ownership. This allows the collaborative governance model to simulate a shared data environment while maintaining compliance with privacy requirements.

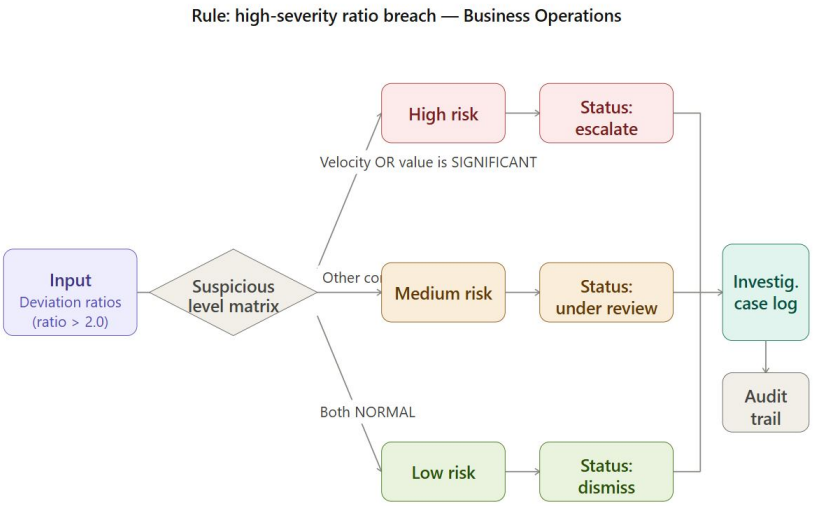

Figure 5. Business Operations triage and escalation workflow for high-severity ratio breach. Cases flagged by the data operations pipeline are triaged based on suspiciousness label: High-risk cases are escalated to the second line, Medium-risk cases are placed under review with an investigation case log, and Low-risk cases are dismissed, with all outcomes documented in the audit trail.

Thus, synthetic data generation is treated as a governance infrastructure component that enables experimentation with alternative data access architectures. Still, it is not a replacement for real data and carries inherent risks — including inherited bias from source data, distributional drift over time, and potential privacy vulnerabilities such as membership inference. In response, the author acknowledges this and recommends governing the synthetic data lifecycle to minimize the risks.

4. Experimental Setup and Results

4.1 Experimental Setup

All experiments were conducted in a controlled computational environment to ensure reproducibility and consistency between the segmented and collaborative simulations. The hardware environment consisted of an AMD Ryzen 5 3550H processor (2.10 GHz), an NVIDIA GeForce GTX 1050 GPU with CUDA support (used for synthetic data generation), 16 GB of RAM, and SSD storage for data handling.

The software stack was built on Python 3.12.3. Data manipulation and aggregation were performed using Pandas 3.0.1 and NumPy 1.26.4. Scikit-learn 1.8.0 provided data scaling, label encoding, and train-test splitting. Synthetic data generation used the TabDDPM (diffusion) model implemented within the SynthCity framework [31], which also supplied built-in quality evaluation metrics for fidelity, utility, and privacy. Independent external validation was performed using SynthEval [32], which adds membership-inference attack simulation and attribute disclosure risk assessment. Distribution visualizations were produced using Matplotlib 3.10.8 and Seaborn 0.13.2 during development, but are not reported in this section. Additionally, a fixed random seed of 42 was used throughout sampling, model training, and generation to ensure full reproducibility across both simulation scenarios.

The experimental workflow proceeded in five stages. First, the original banking data was loaded, and stratified sampling was applied using Pandas, producing a representative baseline of 100 customers with 16,797 transaction records. This baseline dataset served as the common starting point for both simulation tracks. Second, the segmented governance (AS-IS) simulation was executed directly on the real baseline dataset, emulating IT-mediated access with formal requests and information silos across the three lines of defense.

Third, synthetic data was generated from the same baseline using the TabDDPM model within SynthCity [31]. Quality validation was performed using two approaches: SynthCity's built-in evaluation metrics for fidelity, utility, and privacy; and an independent external validation using SynthEval [32], before proceeding to the collaborative simulation (TO-BE).

Fourth, the collaborative governance (TO-BE) simulation was executed using the validated synthetic dataset, assuming a shared, direct-access environment across all governance lines. Fifth, a comparative analysis was performed using custom Python scripts that computed governance

metrics, including Granularity Compression Ratio (Granularity Compression Ratio (GCR)), handoff counts, and Reconstruction Dependency Ratio (Reconstruction Dependency Ratio (RDR)), across both simulations. All random seeds were fixed to ensure that identical monitoring and escalation logic was applied through the same Python functions in both scenarios.

4.2 Dataset and Sampling Outcomes

The study uses a banking transaction dataset containing customer demographics and transaction history. The original dataset includes customer records with attributes such as age, tenure, gender, language preference, and transaction details, including amount, currency, and timestamps.

To ensure representativeness while managing computational constraints, a stratified sampling strategy was applied based on customer transaction activity levels. Four activity categories were defined: inactive/dormant (fewer than 12 transactions per year), low (12–50 transactions per year), moderate (50–200 transactions per year), and severe (more than 200 transactions per year). The sampling was designed to maintain the population distribution of each category, as shown in Table 4.1.

Table 2. Stratified Sampling Outcomes

Activity Level	**Criteria**	**Population (%)**	**Sample (%)**	**Difference**
Inactive/Dormant	< 12 transactions/year	6.4	6.0	0.4
Low	12–50 transactions/year	12.3	12.0	0.3
Moderate	50–200 transactions/year	60.6	61.0	0.4
Severe	> 200 transactions/year	20.7	21.0	0.3

The final experimental dataset comprises 100 customers with complete transaction histories, totaling 16,797 transaction records. All simulations (both segmented and collaborative) use this dataset as the baseline reference. The simulation period covers January 2024 to December 2024.

Tables 3 and 4 reflect a deliberate methodological choice: both simulations use the same underlying dataset to ensure that any observed differences in outcomes are attributable solely to the governance architecture, not to data variation. This single-variable design is necessary because the study does not aim to optimize detection performance but to demonstrate whether synthetic data can serve as a viable governance layer that enables proactive, cross-functional

risk identification — particularly for functions such as internal audit, which traditionally operate reactively on sampled or delayed data [2, 4].

The choice to use real data in the AS-IS simulation and synthetic data in the TO-BE simulation is not arbitrary. Real data is used in AS-IS precisely because information silos in practice emerge from the sensitivity of real operational data – organizations restrict access not due to technical limitations but due to legitimate data protection and governance requirements [1, 3]. Resolving this tension by simply expanding access to real data would introduce unacceptable custodianship risks and violate GDPR data minimization principles. Synthetic data in TO-BE, therefore, represents a structural solution rather than a technical workaround: it separates analytical access from data sensitivity, enabling all governance lines to work with statistically equivalent information without exposing personally identifiable records.

Table 3. Simulation Framework Comparison

Aspect	**AS-IS Simulation**	**TO-BE Simulation**
Framework	Three Lines of Defence (3LoD)	Three Lines Model (3LM) — IIA 2020
Data type	Real data — PII present	Synthetic data — no PII
Data sharing	Siloed — each line sees its own view	All lines have direct access (Golden Source)
Monitoring logic	Identical deterministic rules	Identical deterministic rules
Input customers	100 sampled customers	100 synthetic customers
Input transactions	∼16,797 rows	∼16,797 synthetic rows

Table 4. Post-Simulation Dataset Overview

Metric	AS-IS	TO-BE (SDG)	Change
Dataset used	Original Estonia Bank data	Synthetic data (DDPM via SynthCity)	Different source, same structure
Customer records	100 customers	100 customers	Identical count
Transaction records	16,797 transactions	16,797 transactions	Identical count
Customer-day records	11,717	11,159	−558 records (synthetic compression)
Simulation period	Jan 2024 – Dec 2024	Jan 2024 – Dec 2024	Identical period
Customer identifier	Client ID (real PII)	Pseudo Customer Key (anonymous)	PII removed in TO-BE
Data access model	Request-based via IT	Direct shared access (Golden Source)	Fundamental architecture change

4.3 Evaluation Metrics and Baseline Definition

To compare the segmented (3LoD) and collaborative (3LM) governance architectures, a set of performance metrics is defined. These metrics focus on governance efficiency, data accessibility, and structural friction – not on predictive accuracy.

The segmented (3LoD) model serves as the baseline. All improvements are measured relative to this baseline, which represents current industry practice where data access is IT-mediated, permission-based, and siloed. Three categories of metrics are used, each aligned with the conceptual dimensions defined in section 3.

Information granularity and abstraction metrics include the (GCR), which is the ratio of analytical observations (customer-day aggregates) to raw transaction records; a lower value indicates greater abstraction; and similarity between architectures confirms consistent aggregation logic. The Field Reduction Ratio (Field Reduction Ratio (FRR)) measures the proportion of data attributes retained as information moves from raw data to analytical views; higher values indicate less fragmentation. The Visibility Transformation Index is a qualitative assessment of whether demographic and linkage information is available to lines 2 and 3 (risk management, compliance, and audit).

Escalation and workload metrics include the escalation rate, which is the proportion of monitored cases (customer–days) that trigger a formal alert; comparable rates across architectures confirm consistent detection logic. The amplification ratio measures the volume of transaction records retrieved per escalated case; higher values indicate deeper analytical capability. Access share is the percentage of the total transaction population that can be explored autonomously by a governance line without IT mediation.

Structural dependency metrics include handoff count, the number of inter-departmental interactions required to resolve a single escalated case; lower counts indicate reduced coordination overhead. The Reconstruction Dependency Ratio (RDR) is the percentage of escalated cases that require IT-mediated data extraction; lower values indicate greater analytical autonomy. The escalation dependency loop is a binary indicator of whether access after a red flag requires sequential handoffs (segmented) or is concurrent and direct (collaborative).

Governance quality metrics (control variables) include SOP compliance rate, escalation integrity rate, and documentation sufficiency rate. These metrics are expected to produce 100% in both simulations by design: since the simulation applies deterministic rule-based logic, any case that meets the escalation threshold will automatically satisfy the corresponding SOP, documentation, and integrity requirements. These metrics, therefore, serve as internal consistency checks rather than empirical outcomes, confirming that the deterministic pipeline executed without error across both governance architectures.

4.4 Evaluation Logic

The evaluation logic is designed to compare segmented and collaborative governance structures across consistent system-level criteria.

Evaluation is structured along three dimensions: fidelity, utility, and privacy. Fidelity measures the statistical similarity between synthetic and real data distributions. Utility measures how effectively the data supports downstream analytical tasks. Privacy measures the risk of re-identification or the leakage of sensitive information.

Importantly, the evaluation framework is not used to optimize or tune models, but to validate whether the synthetic data layer is suitable for enabling meaningful governance comparison.

The outcome of this evaluation supports the comparative analysis presented in the subsequent sections, where system-level differences between segmented and collaborative governance are quantified.

4.5 Data Quality Validation

Synthetic data quality was evaluated across three dimensions: fidelity (statistical realism), utility (analytical usefulness), and privacy (protection against re-identification). Two tools were used: SynthCity built-in evaluation and SynthEval for independent validation.

4.5.1 Customer Data Validation

For customer data, the validation results are summarized in Table 4.2. The data-mismatch sanity score was perfect (0.0), confirming that there are no schema inconsistencies between the real and synthetic data. The common rows proportion was 0.11, indicating that only 11% of rows overlapped; this is safe and shows the synthetic data did not simply copy the original. The Jensen–Shannon divergence was 0.0098 (very good), and the KS test marginal was 0.935 (also very good), indicating that the synthetic distribution is statistically equivalent to the real data. Mean differences were small: age differed by 0.7% (real 43.46 years vs synthetic 43.75 years), and tenure differed by 8.2% (real 5.95 years vs synthetic 5.46 years), both within acceptable tolerance for governance simulation.

Table 5. Synthetic Customer Data — Fidelity and Privacy

Metric	Score	Interpretation
Data Mismatch (Sanity)	0.0 (Perfect)	No schema inconsistencies.
Common Rows Proportion	0.11 (Safe)	Only 11% row overlap — no simple copying.
Jensen-Shannon Divergence (Jensen-Shannon Divergence (JSD))	0.0098 (Very Good)	Statistical distributions closely match real data.
KS Test Marginal (Kolmogorov-Smirnov (KS))	0.935 (Very Good)	Distribution statistically equivalent to real data.
Mean Difference: Age	0.7% (Very Good)	Real 43.46 vs synthetic 43.75 years.
Mean Difference: Tenure	8.2% (Good)	Real 5.95 vs synthetic 5.46 years.

4.5.2 Transaction Data Validation

The results of transaction data validation are presented in Table 4.3. The data mismatch was perfect (0.0), and the proportion of common rows was 0.0 – no rows were repeated from the original, indicating full privacy preservation. The Jensen–Shannon divergence was 0.00192 (excellent), and the KS test marginal was 0.985 (excellent), showing that the synthetic distributions are nearly identical to real data. The ML utility ratio was 0.997, meaning that a machine learning model trained on synthetic data achieves 99.7% of the performance of one trained on real data – an excellent result for downstream analytical tasks. Mean differences for transaction amount were 4.7% for amount in euros (real €160.28 vs synthetic €152.68) and 11.2% for the original amount (real €175.80 vs synthetic €156.04); both are within acceptable tolerance given the skewness of financial data. The identifiability score was below the safe threshold, indicating a very low risk of re-identifying real customers from the synthetic data. Training time for the DDPM was 13.3 minutes, which is efficient for 1,000 epochs on a GPU.

Table 6. Synthetic Transaction Data — Fidelity, Utility, and Privacy

Metric	Score	Interpretation
Data Mismatch (Sanity)	0.0 (Perfect)	Identical schema, no type inconsistencies.
Common Rows Proportion	0.0 (Perfect)	No rows repeated from original — full privacy.
Jensen-Shannon Divergence (JSD)	0.00192 (Excellent)	Distributions nearly identical to real data.
KS Test Marginal	0.985 (Excellent)	Almost perfect distribution match.
ML Utility Ratio	0.997 (Excellent)	ML model performs at 99.7% of real-data performance.
Mean Difference: Amount (EUR)	4.7% (Good)	Real €160.28 vs synthetic €152.68.
Mean Difference: Amount (Original)	11.2% (Acceptable)	Real €175.80 vs synthetic €156.04.
Identifiability Score	Below safe threshold	Very low risk of re-identifying real customers.
DDPM Training Time	13.3 minutes	Efficient for 1,000 epochs on GPU.

4.5.3 Summary of Synthetic Data Quality

Table 7 summarizes the three pillars of synthetic data quality as assessed by the SynthEval framework. Fidelity (statistical faithfulness) was excellent: all Jensen–Shannon distances were below 0.002, KS test values exceeded 0.93, and mean differences for most features were under

5%. Utility (analytical usefulness) was also excellent, with an ML utility ratio of 0.997 and consistent anomaly detection patterns between real and synthetic data. Privacy (data protection) was excellent: common rows were 0% for transactions, the identifiability score remained below the safe threshold, and all PII was removed before model training. These results confirm that the synthetic data is statistically faithful, analytically useful, and privacy-preserving, making it suitable for governance simulation and comparative analysis.

Table 7. Three Pillars of SDG Quality — SynthEval Summary

Dimension	Result
Fidelity — statistical faithfulness	Excellent: JSD < 0.002, KS test > 0.93, mean differences $< 5\%$ for most features.
Utility — analytical usefulness	Excellent: ML utility ratio 0.997 (99.7% of real data), anomaly detection patterns consistent.
Privacy — data protection	Excellent: common rows 0% (transactions), identifiability score below threshold, PII removed before training.

4.6 Segmented (AS-IS) Simulation Execution

The segmented model (3LoD) simulation implements the traditional IT-mediated, permission-based data access process. The workflow follows a sequential chain: monitoring, then escalation, then a formal data request to IT, then a processing delay, then aggregated data delivery, and finally audit reconstruction. This process was implemented using custom Python scripts that simulated IT request queues and response times based on the actual number of escalated cases.

The choice to measure IT processing lag as a primary bottleneck indicator reflects the operational reality of governance functions operating under fixed project timelines. In practice, delayed data access does not merely slow investigations—it fundamentally limits the scope and relevance of audit findings. When data requests remain unresolved within an audit cycle, functions are forced to either proceed with incomplete information or classify cases as unauditable, reducing assurance activities to procedural compliance checks rather than substantive risk assessments [2, 11].

The average IT processing lag of 0.6 days per request was derived from a priority-based Service Level Agreement (SLA) framework, informed by Control Objectives for Information and Related

Technologies (COBIT) 2019 Deliver, Service, and Support-Managed Service Requests and Incidents (DSS02) and Information Technology Infrastructure Library (ITIL) 4 service request management practices. Three priority tiers were defined based on the suspiciousness label: Urgent (high cases, fulfillment in < 24 hours), High (medium cases, 1–2 days), and Normal (low cases, 3–5 days). The resulting weighted average of 0.6 days reflects the distribution of 1,806 escalated cases across priority tiers, with the majority classified as High or Urgent due to the nature of suspicious transaction monitoring.

Key outcomes of the segmented simulation are as follows. Total monitoring observations (customer—days) were 11,717, resulting in 1,806 escalated cases, yielding an escalation rate of 15.41%. Each escalated case generated 1 formal IT request, for a total of 1,806 requests. The average IT processing lag was 0.6 days per request. During request handling, 6,319 rows of personally identifiable information (PII) were exposed to IT personnel and to the requesting personnel. The amplification ratio – the number of transaction records retrieved per escalated case – was 2.03×. The handoff count, measuring inter-departmental interactions, was 3 per case (from DataOps to Business Ops, then to Risk/Compliance, and finally to Audit). The Reconstruction Dependency Ratio (RDR) was 100%, meaning every escalated case required IT-mediated data extraction. Cross□line data join was impossible because each line operated on fragmented views, and demographic information was not accessible to lines 2 and 3. Fragmentation effects included incomplete audit populations (sample–based only) and duplicated testing across compliance and risk functions.

4.7 Collaborative (TO-BE) Simulation Execution

The collaborative model (3LM) simulation replaces IT-mediated access with a shared synthetic data environment. All governance lines access the same synthetic customer and transaction data directly, without permission requests for routine monitoring and analysis. The same escalation logic (deviation ratios, grouping, labeling) was applied using identical Python functions, ensuring comparability.

Key outcomes of the collaborative simulation are as follows. Total monitoring observations were 11,159 customer-days – a similar scale to the segmented simulation. Escalated cases numbered 1,709, yielding an escalation rate of 15.31%, which is virtually identical to the segmented rate (15.41%), confirming that the detection logic produced consistent behavioral signals on synthetic data. IT requests generated for routine analysis were zero. Investigation lag was

zero days because access was real-time and direct. No PII rows were exposed because the synthetic data contains no real personally identifiable information. The amplification ratio was 2.54× higher than in the segmented simulation, because all historical transaction records became instantly accessible. The handoff count was zero because there were no inter-departmental data-provisioning steps. The Reconstruction Dependency Ratio (RDR) was 0%, because all drill-downs were self-serve. Cross□line data join was fully enabled via the pseudo□customer key, and demographic information was accessible to lines 2 and 3, having been enriched in the synthetic golden source. Parallel access across lines enabled concurrent investigation without sequential delays. Specific IT requests occurred only when data was needed as formal evidence for legal or compliance purposes, not during routine daily analysis.

4.8 Governance Simulation Results Matrices

This section presents three performance matrices comparing the segmented and collaborative governance architectures across the dimensions defined in Section 3.2.

4.8.1 Matrix 1 — Information Granularity & Abstraction

This matrix shows how raw event-level data is compressed and abstracted throughout the monitoring pipeline, illustrating changes in information visibility across different functions.

Table 8. Data Layer Visibility: AS-IS vs TO-BE

Layer	Primary Dataset	Fields	AS-IS Access	TO-BE Access
IT	Transactions Sampled / Synthetic Transactions	13 / 16	IT Only	Produces SDG
DataOps	Customer Day / Synthetic Customer Day	14 / 18	DataOps → BizOps only	All lines (Golden Source)
Business Ops	Ops Case Log / TO-BE Ops Case Log	12	BizOps only	All lines
Risk Mgmt	Risk Dashboard / TO-BE Risk Dashboard	13	2LoD only (aggregated)	All lines (granular + aggregate)
Compliance	Compliance Log / TO-BE Compliance Log	12	2LoD only	All lines
Internal Audit	Audit Review Log / TO-BE Audit Review Log	25+	Request-based (all silos)	Direct (no IT request)

Table 9. Information Granularity & Abstraction Matrix

Metric	AS-IS (3LoD)	TO-BE (3LM)	Change
Total transactions (input)	∼16,797	∼16,797	= Same scale
Total customer-days produced	11,717	11,159	= Similar scale
Granularity compression ratio	0.70	0.66	= Similar compression
Demographics at 2/3LoD	No	Yes	**New capability enabled**
Cross-line data join	Not possible	Enabled	**Pseudo Customer Key linkage**
PII in monitoring data	Yes	None	**Replaced by Pseudo Customer Key**

Both simulations use the same method of operating on transaction-level data down to the customer-by-day level and produce similar compression ratios (∼0.66–0.70). The structural difference lies in what happens after compression: in AS-IS, each row receives only its own abstract view with no cross-row join capability. In TO-BE, all rows access the same Primary Source simultaneously, and demographic data is enriched during the data engineering process in Data Operations – enabling Risk and Audit to cross-reference customer profiles without any IT requests.

4.8.2 Matrix 2 — Operational Escalation & Workload

This matrix illustrates the level of process escalation and workload to evaluate monitoring selectivity, IT dependency, and the operational burden generated per escalated case.

Table 10. Operational Escalation & Workload Matrix

Metric	AS-IS (3LoD)	TO-BE (3LM)	Change
Total monitoring observations	11,717	11,159	∼Equivalent
Total escalated cases	1,806 (15.41%)	1,709 (15.31%)	Consistent rate
Txn records (escalated)	3,658 records	4,343 records	+685 more accessible
Amplification ratio	2.03×	2.54×	More depth in TO-BE
IT requests generated	1,806	0	**Fully eliminated**
Avg IT processing lag	0.6 days per case	0 days	**Eliminated**
PII rows exposed	6,319	0	**Zero exposure**
Self-serve drill-down	No	Yes	**Immediate — no IT needed**

The escalation rates for both simulations were nearly identical (15.41% vs 15.31%)—confirming that the monitoring logic produced equivalent detection behavior across real and synthetic data. The post-escalation process in AS-IS generated 1,806 data requests to IT, exposing 6,319 rows PII and causing an average delay of 0.6 days. TO-BE eliminated all three–zero requests, zero PII exposure, and zero delay.

The higher amplification ratio in the TO-BE simulation (2.54× vs 2.03×) reflects an improvement in investigative depth rather than an increase in workload burden. In the AS-IS simulation, the amplification ratio is constrained by the IT request cycle — investigators can only retrieve the specific records they explicitly requested, and any discovered relevance to additional transactions requires a new request cycle. This creates a recursive dependency loop in which the investigative scope is limited not by analytical judgment but by access architecture.

In the TO-BE simulation, all governance lines can autonomously drill down into the full transaction history of any flagged customer without initiating additional requests. The higher amplification ratio, therefore, indicates that investigators can trace suspicious patterns more thoroughly—following transaction linkages to their sources rather than stopping at the boundary of what was pre-approved for retrieval. While this increases the analytical workload per case, it enables immediate execution without the delays and iterative dependencies inherent in the AS-IS request cycle [2, 4].

4.8.3 Matrix 3 — Structural Bottleneck Indicators

This matrix evaluates the structural dependencies and governance constraints built into each architecture.

Table 11. Structural Bottleneck Indicators Matrix

Metric	AS-IS (3LoD)	TO-BE (3LM)	Change
IT handoffs per escalation	3 handoffs	0 handoffs	**−3 (eliminated)**
Lines with direct data access	1 (IT only)	4 (all lines)	**+3 lines unlocked**
Cross-line join possible	Not possible	Fully enabled	**New capability**
IT requests for drill-down	1,806 requests	0 requests	**−1,806 (eliminated)**
PII exposure events	6,319 rows (high risk)	0 rows	**−6,319 rows**
Data access latency	0.6 days avg	0 days (real-time)	**Fully eliminated**
Control effectiveness rate	100.00% (escalated only)	100.00% (escalated only)	Consistent
Escalation integrity rate	100.00%	100.00%	Consistent
Documentation sufficiency rate	100.00%	100.00%	Consistent
SOP compliance rate	100%	100%	= Identical — same rules
Ratio/flag consistency	N/A (real data)	100%	**Deterministic pipeline**

Identical control effectiveness, integrity, and compliance metrics confirm that the predetermined design decisions are valid and both simulations apply the same deterministic logic. The differences in this comparison are entirely due to the data access architecture, not the quality of the controls. The TO-BE simulation eliminates all handoffs, requests, and PII exposure by replacing the IT-mediated loop with a shared synthetic data layer accessible to all lines simultaneously while maintaining governance quality at identical levels.

4.9 Synthetic Data Compression and Analytical Scope

The difference of 558 customer-day records between the AS-IS (11,717) and TO-BE (11,159) simulations is an expected outcome of synthetic data generation rather than a methodological

inconsistency. While the total number of customers and transactions is identical across both simulations, the daily distribution of transactions per customer differs in the synthetic dataset — a natural consequence of statistical reproduction rather than exact replication. This difference is acceptable for the analytical purposes of this study, as the synthetic data maintains near-identical distributional properties ($JSD < 0.002, KStest > 0.93$) and produces a consistent escalation rate (15.31% vs 15.41%), confirming that behavioral risk signals are preserved across both representations.

It is important to note, however, that this distributional equivalence is sufficient for proactive risk monitoring and governance analysis, but not for formal legal or forensic purposes. Synthetic data cannot substitute for real transaction records as legal evidence — its role in this study is explicitly limited to enabling routine analytical access across governance lines, while raw data remains accessible through formally justified, evidence-level requests [1, 4].

4.10 Comparative Analysis

Tables 12 and 13 present the side-by-side comparison of segmented versus collaborative governance across all performance dimensions.

Table 12. Comparative Results — Data Access, Privacy, and Governance Quality

Category	**Metric**	**Segmented (3LoD)**	**Collaborative (3LM+SDG)**	**Change**
Data Access	IT requests (routine)	1,806	0	**−100% (Eliminated)**
	Lines with direct access	1 (IT only)	4 (all lines)	+3 lines unlocked
	Cross-line data join	Not possible	Enabled	New capability
	Average IT/access lag	0.6 days	0 days	**Latency eliminated**
Privacy	PII rows exposed	6,319	0	**−100% (Zero risk)**
	Re-identification risk	High (raw PII)	Below safe threshold	Risk mitigated
Governance	SOP compliance	100%	100%	Consistent
	Escalation integrity	100%	100%	Consistent
	Documentation rate	100%	100%	Consistent

Table 13. Comparative Results — Information Granularity, Operational Escalation, and Structural Dependencies

Category	**Metric**	**Segmented (3LoD)**	**Collaborative (3LM +SDG)**	**Change**
Granularity	GCR	0.70	0.66	Similar compression
	Demographics at lines 2/3	No	Yes	New capability
Escalation	Escalation rate	15.41%	15.31%	Consistent
	Amplification ratio	2.03×	2.54×	+25% more depth
	Access share	Limited (via requests)	Full (autonomous)	Substantial gain
Structural	Handoff count per case	3	0	**−3 eliminated**
	RDR	100%	0%	**−100%**
	Escalation dependency loop	Sequential	Concurrent	Eliminated

Interpretation of the key differences is straightforward. The collaborative model eliminates all IT-mediated bottlenecks: 1,806 requests become 0, 6,319 rows PII become 0, and 0.6 days of lag become 0. Governance quality metrics (SOP compliance, escalation integrity, and documentation rate) remain at 100% in both models by design, as both simulations apply identical deterministic rules–confirming that improved accessibility does not compromise governance standards. The escalation rate is nearly identical (15.41% vs 15.31%), proving that the same behavioral detection logic works consistently on synthetic data as on real data. The higher amplification ratio in the collaborative model (2.54× vs 2.03×) reflects that investigators can instantly access complete transaction histories rather than being limited by IT processing backlogs.

4.11 Summary of Results

The simulation results collectively demonstrate that a synthetic-data-enabled governance architecture can meaningfully address the trade-off between information access and data protection identified in this study, though with important caveats that reflect the complexity of real-world implementation of the organizational governance.

Regarding **SQ1** — whether synthetic data can eliminate silos and reduce reliance on IT handoffs — the simulation provides clear evidence that this is technically feasible. All 1,806 routine IT

requests were eliminated, handoff count reduced from 3 to 0 per case, and investigation lag dropped from 0.6 days to real-time access. However, the author acknowledges that technical feasibility alone does not guarantee organizational transformation. Governance structures are not merely technical systems — they are embedded in organizational culture, maturity, and board-level commitment to change. The most significant barrier to adopting synthetic data as a governance enabler is therefore not technological but structural: organizations must be willing to redefine data ownership boundaries and invest in the governance infrastructure required to sustain a shared synthetic data layer [1, 4].

Regarding **SQ2** — whether data quality is maintained or enhanced — the synthetic data demonstrated strong statistical fidelity (JSD < 0.002, KS test > 0.93) and analytical utility (Machine Learning (ML) utility ratio 0.997), confirming that governance functions can operate on synthetic data without material loss of analytical capability. However, synthetic data quality is fundamentally bounded by the quality of its source data. No generative model can compensate for poor input data — human judgment and continuous oversight of data inputs remain essential components of any synthetic data pipeline. Technology enables the process, but humans must govern it [3].

Regarding **SQ3** — whether compliance and escalation integrity are preserved — both simulations produced identical governance quality metrics: 100% Standard Operational Procedure (SOP) compliance, escalation integrity, and documentation sufficiency under deterministic simulation conditions. This confirms that shifting to a synthetic data architecture does not compromise control effectiveness. However, these results assume that the underlying governance rules are correctly defined and up to date. In practice, rules must be periodically reviewed and updated to reflect regulatory changes — a static rule set applied to a dynamic regulatory environment will eventually produce compliance gaps regardless of the data architecture [1, 2].

The collaborative (3LM+ synthetic data) simulation delivers consistent or superior results across all governance metrics compared to the segmented (3LoD) simulation. Three improvements are most critical. First, 1,806 IT requests are fully eliminated, reducing investigation lag from 0.6 days to zero. Second, 6,319 rows of PII are eliminated, resolving the highest–severity privacy risk. Third, cross-line data access expands beyond a single line (IT only) to all four governance lines simultaneously, enabling collaboration that was structurally impossible under the segmented model.

All governance quality metrics — SOP compliance, escalation integrity, and documentation rate — remain at 100% in both simulations by design, as both simulations apply identical deterministic rules, confirming that improved accessibility does not compromise governance standards. The synthetic data quality validation, with an ML utility ratio of 0.997, confirms that the synthetic data is a reliable analytical foundation for GRC decision-making.

Thus, the answer to the research question is clear: the collaborative governance model, enabled by synthetic data, successfully resolves the governance-privacy paradox. It provides all three lines with simultaneous, privacy-safe, analytically useful data access – eliminating the structural bottlenecks and information asymmetries inherent in the traditional segmented (3LoD) model, while maintaining full governance quality.

However, it is important to note that this study shows a positive result under certain conditions. Specifically, while (1) synthetic data quality is empirically validated through statistical testing, (2) the governance improvements are simulation-based architectural outcomes rather than real-world organizational evidence.

5. Discussion

5.1 Answering the Research Questions

The governance challenges documented in this study are not abstract organizational phenomena — they are experienced daily by practitioners navigating competing demands of compliance, data protection, and operational efficiency. Structural inefficiencies such as delayed data requests, fragmented audit trails, and asymmetric information flows are well-documented in the literature [2, 3], yet persist in practice precisely because addressing them requires the kind of sustained attention that high-workload operational environments rarely afford [4, 11]. This study approaches the trade-off between information access and data protection from the practitioner perspective—not only as a technical problem to be solved, but also as a structural condition to be understood and, ultimately, a challenge that requires collaborative effort across functions, disciplines, and organizational boundaries to be meaningfully addressed.

SQ1: Can synthetic data eliminate silos and reduce reliance on IT handoffs?

Technically, yes. The simulation eliminated all 1,806 routine IT requests, reduced handoffs from 3 to 0 per case, and eliminated investigation lag entirely. But the more honest answer is: it depends on whether organizations are ready to let it. The tension between data utility and compliance is not primarily a data problem—it is an organizational one. Functions that are already stretched thin rarely have the capacity to champion structural change, even when they recognize its value. Innovation requires incentive structures, board-level commitment, and the willingness to absorb short-term effort for long-term gain [1, 2]. This study demonstrates that the technical barrier is feasible. The harder question — why organizations with access to both governance expertise and technical capability still operate in silos — points to something more fundamental: the cost of thinking differently when thinking itself is a scarce resource.

SQ2: Does the transition improve data quality while maintaining analytical completeness?

The synthetic data performed well across all quality dimensions — JSD < 0.002, KS test > 0.93, and an ML utility ratio of 0.997. In simulation terms, this is a strong result. In practice, it is a starting point. Banking transaction data carries years of accumulated complexity —input errors, system logic inconsistencies, core banking migrations, and institutional mergers that can take years to stabilize. No generative model can fully reproduce this messiness, and none should be expected to [3]. Synthetic data quality is ultimately bounded by the quality of its source —

which means that the human work of maintaining clean, consistent, well-governed operational data is not replaced by this approach, but made more consequential. Garbage in, garbage out — just faster and at a greater scale.

SQ3: Can decentralized access fully preserve compliance and escalation integrity?

In the simulation, yes — 100% across all governance quality metrics. But this result should be read carefully. The simulation is deterministic by design: any case that meets the escalation threshold automatically satisfies all corresponding governance requirements. Real compliance environments do not work this way. Rules have gaps, judgment varies, and regulatory frameworks evolve faster than internal controls are updated [1, 2]. The 100% result demonstrates that synthetic data does not introduce additional governance risk — not that it eliminates the risk that already exists. Compliance, in practice, is not a destination but an ongoing negotiation between what the rules say, what the data shows, and what the organization is actually capable of doing. This study contributes one piece of that puzzle — and hopes to prompt others to contribute the rest, together.

5.2 Governance Shift: From Segmented to Collaborative Models

The results demonstrate that moving from a segmented (3LoD) governance architecture to a collaborative (3LM+ synthetic data) architecture fundamentally alters how GRC functions, accesses, and shares information. The segmented model operates as a request-driven hierarchy: each line depends on IT for data, creating sequential dependencies, delays, and PII exposure. The collaborative model operates as an access-enabled network: all lines connect directly to a shared synthetic data layer. The elimination of 1,806 IT requests, reduction of lag from 0.6 days to zero, and removal of 6,319 PII rows from routine workflows are not incremental improvements – they represent a structural transformation of the governance information system. The near-identical escalation rates (15.41% vs 15.31%) confirm that this transformation does not alter detection logic; it only improves speed, autonomy, and privacy.

5.3 Synthetic Data as a Structural Enabler

Synthetic data functions as a governance abstraction layer, not merely a privacy-preserving technique. By providing a statistically equivalent, PII–free representation of operational data, it decouples analytical work from raw data sources. This decoupling resolves the friction between transparency and confidentiality: organizations can implement the collaborative intent of the

3LM without violating GDPR data minimization requirements. The high fidelity (JSD < 0.002, KS test > 0.93), utility (ML ratio 0.997), and low identifiability risk validate that synthetic data is fit for governance purposes. The shift from sequential handoffs (3 per case) to parallel investigation (zero handoffs) and from 100% reconstruction dependency to 0% demonstrates that synthetic data enables analytical autonomy for all lines.

5.4 Governance Implications and Trade–off Resolution

The findings have theoretical and practical implications. Theoretically, governance models should be evaluated as information architectures – the data access design matters as much as role definitions. Synthetic data shifts from a technical tool to a governance enabler, and behavioral deviation becomes a consistent analytical construct across real and synthetic domains. Practically, GRC functions can achieve real-time, collaborative assurance without compromising privacy. The trade-off between collaboration and data protection is resolved not by weakening either, but by substituting a validated synthetic layer for raw data in routine workflows. Raw data access is preserved for legal evidence but becomes the exception. Future work should test generalisability across use cases, scales, and regulatory contexts, and examine practitioner trust and regulatory acceptance.

These findings contribute to the theoretical development of practical expectations from updating governance models from the 3LoD to the 3LM — improving data accessibility, quality, and integrated assurance and decision-making purposes, while reducing siloed systems and increasing data relevance. A synthetic data-enabled architecture can directly address the core barriers of the 3LoD model — fragmented data, siloed functions, and lack of cross-line integration — by supporting the 3LM 's emphasis on collaboration, flexibility, and accountability over defensiveness alone, enabling secure and privacy-preserving downstream data access across all three lines without relying on rigid, reactive structures that hinder rapid adaptation to changes in the business environment or technology.

5.5 Organizational Readiness and Institutional Trust

Even if the technical solution is available, the hardest part of implementing a new approach is overcoming structural resistance — and this resistance operates at two levels.

At the organizational level The barrier is not irrational: it reflects the legitimate cost of change in high-stakes environments, where incentive misalignment, operational instability, and capacity

constraints converge. Functions that are already stretched thin rarely have the capacity to champion structural transformation, even when they recognize its value — and this is not incidental. The uneven distribution of operational burden is itself a structural consequence of siloed governance: functions closest to the data carry the heaviest workload, yet operate with the least cross-functional visibility. As a result, those most affected by the inefficiency are least positioned to change it.

At the institutional level The barrier is equally significant: highly regulated industries require regulatory legitimacy before any transformation in GRC workflows can be meaningfully endorsed. There remains a gap between technical validity and institutional legitimacy that is perhaps most visible in audit practice itself, where confidentiality norms are deeply embedded. Auditors are trained to rely on real evidence, and establishing synthetic data as a credible governance instrument requires sustained proof-of-concept efforts that go beyond statistical validation — it requires trust, which is built incrementally and cannot be mandated.

5.6 Relation to Prior Work

This study sits at the intersection of two research streams that have, until now, developed largely in parallel, without meaningful exchange or integration. The separation is not incidental—it reflects a deeper disciplinary boundary between synthetic data research, which is predominantly situated in computer science and statistics, and governance assurance research, which draws on accounting, internal audit, and regulatory studies. Researchers in each stream cite different journals, address different audiences, and frame problems in fundamentally different ways. This study argues that the boundary itself is part of the problem it seeks to address.

Within the GRC and internal audit literature, studies consistently document the structural limitations of siloed governance — fragmented data access, reactive assurance processes, and the persistent gap between the collaborative intent of the 3LM and its operational reality [1–3]. Deloitte and McKinsey similarly highlight that integrated assurance remains an aspiration rather than a standard practice in most organisations [4, 11]. However, this literature does not provide a technically viable mechanism for resolving the governance–privacy paradox — it identifies the problem without offering a structural solution that satisfies both collaboration and data protection requirements simultaneously.

Within the SDG literature, research has demonstrated strong technical capability for privacy-preserving analytics across healthcare, finance, and cybersecurity [12, 20]. Reite et al. [23] identified the type of dynamic, behavioral data needed for effective suspicious transaction monitoring; Cardoso et al. [30] demonstrated the structural value of graph-based transaction representations; and Chen et al. [29] showed how synthetic generation can address class imbalance in AML detection. Yet none of these studies position synthetic data as a governance infrastructure layer — their objective remains improving ML performance within isolated analytical workflows, not enabling cross-functional assurance.

This study bridges that gap by reframing synthetic data not as a tool for model training, but as a structural enabler of integrated governance. Rather than asking ”how well does synthetic data perform on ML tasks?”, this study asks ”can synthetic data separate analytical access from data sensitivity in a way that makes the 3LM operationally feasible?” The governance metrics used in this study — handoff count, investigation lag, RDR, and GCR — are deliberately distinct from standard SDG evaluation metrics, reflecting the shift from technical performance to governance effectiveness as the primary evaluation criterion.

In doing so, this study makes a contribution that neither stream has made independently: it demonstrates that the collaborative intent of the 3LM can be operationalized in practice, using synthetic data as a shared infrastructure layer that preserves privacy while enabling real-time, self-serve access across all governance lines. The disciplinary separation that prevented this connection earlier is itself a finding—one that points to the value of cross-disciplinary dialogue in addressing governance challenges that are simultaneously technical, organizational, and regulatory.

5.7 Practical Recommendations

1. **Adopt the SDG pipeline as the foundation of the GRC data architecture.** Synthetic data can be used as the foundation of a company’s Governance, Risk, and Compliance (GRC) data architecture. Then begin implementing several activities in stages, for example, focusing on monitoring suspicious transactions as a pilot project first, then moving on to the next stage.

2. **Implement role-based access control (Role-Based Access Control (RBAC)) for the shared data layer.** Implementing a synthetic data architecture with RBAC for the shared

data layer within the organization. Ensuring all three lines function under a governance framework directly accesses synthetic data while remaining governed by clear RBAC policies to maintain proper separation of duties.

3. **Integrate with GDPR, DORA, and relevant regulatory requirements.** The use of synthetic data must comply with the integrated boundaries of the GDPR and DORA, as well as related regulatory requirements. For example, providing documentation of synthetic data use in the Data Protection Impact Assessment (DPIA) to maintain compliance with GDPR Art. 35. Ensuring alignment with the requirements of the DORA, or other regulations, such as the European Banking Authority (EBA).

4. **Build SDG literacy across GRC functions.** Build synthetic data-based literacy across GRC functions to ensure risk management, compliance, and internal audit stakeholders understand how to correctly interpret and apply synthetic data in their respective roles.

5. **Conduct full-scale validation.** Conduct full-scale validation of the synthetic data approach. This requires testing the framework on a comprehensive dataset with adequate infrastructure to assess the performance of synthetic data at scale and in high-complexity problems.

6. **Synthetic data life-cycle governance** Synthetic data resolves the trade-off between information access and data protection, but it also introduces a second layer of governance. However, this second layer is not a separate layer; it is still part of the main solution – positioning synthetic data as a mechanism for seamless governance and integration. Synthetic data life-cycle governance is needed to ensure that the data quality used in the context is accountable without introducing a new approach to the existing governance framework — 3LM.

7. **Ensure reproducibility of the simulation pipeline.** The full simulation pipeline is made publicly available. See the Reproducibility section for the details.

6. Conclusion

To illustrate the problem addressed in this study, consider a typical compliance investigation in a banking environment. A business operations department detects a potentially suspicious transaction pattern and needs to inspect detailed transaction-level data to validate the case. In a traditional Governance, Risk, and Compliance (GRC) setup, this is not straightforward. The analyst must submit a formal request to IT, wait for approval, and receive a filtered dataset. This process introduces delays, additional workload, and unnecessary exposure of sensitive customer data, even when only a small subset of records is required.

This simple example exposes a structural limitation in GRC architectures: although all three lines require access to operational data, privacy regulations such as GDPR and organizational silos prevent direct, real-time access. This creates what is known as the *governance–privacy paradox*: effective governance depends on data access, but privacy constraints restrict it.

This thesis investigated whether synthetic data can resolve this friction by enabling privacy-preserving yet analytically useful data access across all governance functions. Instead of relying on IT-mediated data extraction, we simulate a collaborative Three Lines Model (3LM) in which synthetic data is generated from operational records and shared across business operations, risk management, compliance, and internal audit.

In this synthetic-data-enabled setting, the same investigation described above changes fundamentally. Rather than waiting for IT extraction, the compliance analyst can immediately access a synthetic dataset that preserves statistical and behavioural patterns of the original data while containing no personally identifiable information. Risk, audit, and business functions can simultaneously analyze the same dataset, enabling parallel investigation instead of sequential handoffs.

This study contributes in four ways. Conceptually, it positions synthetic data as a governance infrastructure layer for the Three Lines Model rather than a purely technical tool. Methodologically, it introduces a simulation framework for evaluating governance architectures using metrics such as handoff count, investigation delay, and information consistency. Empirically, it provides evidence that synthetic data can eliminate structural bottlenecks while preserving governance quality. Practically, it offers a blueprint for implementing self-serve, privacy-preserving GRC workflows in regulated environments.

The simulation results show several consistent improvements. All routine IT-mediated data requests were eliminated, removing a major source of operational delay. Investigation time decreased from approximately 0.6 days to real-time access, enabling immediate analysis of suspicious patterns. At the same time, exposure to personally identifiable information was completely eliminated.

These results can be interpreted across four dimensions. First, data access becomes immediate and self-serve rather than request-driven. Second, data quality improves in terms of consistency, since all functions operate on a single shared synthetic representation rather than fragmented extracts. Third, collaboration becomes parallel rather than sequential, enabling simultaneous investigation across governance functions. Fourth, privacy protection is structurally enforced by design, since sensitive data never leaves controlled storage in operational workflows.

Overall, the findings demonstrate that synthetic data fundamentally changes how GRC processes operate. Instead of sequential workflows mediated by IT, organisations can enable real-time, collaborative assurance while preserving privacy constraints. This directly addresses the friction between transparency and confidentiality by decoupling analytical access from exposure to real data. Still, governance effectiveness was preserved under deterministic simulation conditions, indicating that removing data friction did not weaken control outcomes.

6.1 Limitations

However, several limitations must be acknowledged. The evaluation is based on a simulated environment rather than a production deployment, uses a single banking use case, and relies on simplified detection logic. In addition, synthetic data does not replace real data for legal or forensic evidence, and its effectiveness depends on the quality of the generative model and underlying data distributions.

We acknowledge the limitations of simulating processes and generating synthetic data, including:

- In both simulations, suspicious transaction flaggers used a simple deterministic rule-based mechanism, which does not capture real-world noise, process complexity, and human judgment variability. We acknowledge that in real-world contexts, processes and activities within a GRC environment are complex, rigid, and dependent on organizational and juridistic regulatory conditions.

- Relatedly, governance quality metrics — SOP compliance, escalation integrity, and documentation sufficiency rate — return 100% in both simulations by design, as any case meeting the escalation threshold automatically satisfies all corresponding governance requirements under the deterministic pipeline. These metrics, therefore, function as internal consistency checks rather than empirical measures of governance effectiveness. In a real-world deployment, compliance rates would be subject to human judgement, process variability, and organisational factors that deterministic simulation cannot capture.
- Synthetic data can statistically produce data equivalent to real data, but its functionality remains limited in assisting incident analysis and mitigating operational risk. When a real-life case arises, synthetic data cannot replace real data for evidentiary purposes in legal proceedings because it lacks the legal weight of evidence required for formal chain-of-custody purposes.
- The synthetic data pipeline itself carries inherent governance risks — including data quality drift, model misuse, and over-reliance on synthetic outputs in contexts that require real data — which this study does not empirically address.
- The simulations covered only one use case—suspicious transaction monitoring. In a real-world setting, GRC implementations would encompass many additional use cases, such as KYC, AML, and credit risk monitoring, with varying data complexities.
- The sample size used in this study was limited; therefore, performance on the full dataset requires further testing.
- Generalizability of the findings requires validation on a larger and more diverse dataset.

Furthermore, the simulation in this study necessarily simplifies the operational complexity of real banking environments. However, as documented in the GRC literature and in practice, transaction data accumulates inconsistencies arising from system interdependencies, legacy infrastructure, and organizational transitions, which significantly affect data quality and consistency [2, 3].

The deterministic design of the simulation, while methodologically necessary for controlled comparison, produces governance quality metrics of 100% that cannot be expected in real-world deployments, where human judgement variability and regulatory evolution introduce compliance gaps that no rule-based system fully anticipates.

6.2 Future Work

This approach is expected to extend beyond the banking sector. It can also be applied to other highly regulated sectors, such as healthcare and insurance, with systematic adjustments to the applicable regulatory frameworks in each sector. Building on the foundation established in this study, a synthetic data-based workflow can be further developed to assist in:

- **Cross-sector application.** The synthetic data pipeline could be tested in other regulated sectors such as healthcare and insurance, with appropriate adjustments to the applicable regulatory frameworks.
- **Increased data complexity.** Future studies could extend the approach to higher-volume datasets and more complex feature sets, including unstructured data such as documents and transaction notes, to assess performance under more realistic operational conditions.
- **Federated and multi-party settings.** The governance model could be explored in federated learning contexts, where synthetic data enables cross-institutional collaboration without exposing raw operational data across organizational boundaries.
- **Integration with Synthetic Data Governance Frameworks.**

  Future research could examine how synthetic data pipelines should be governed—including how the synthetic data lifecycle is assessed. For example, source data quality, model development, ongoing validation, and user literacy across governance functions are key elements to ensuring accountability. Importantly, synthetic data lifecycle governance requirements do not require the introduction of a separate framework. The governance concept itself—the 3LM—naturally provides an accountability structure that can be adopted across the synthetic data lifecycle. For example, a conceptual walkthrough of synthetic data governance implementation within 3LM: (1) the first line can be positioned as source data quality and integrity layer; (2) the first and second line together develop the synthetic data pipeline and validate it; (3) the second line can be positioned also as evaluator who evaluates and monitors the ongoing process; (4) the third line as an independent assurance for the whole process; and the last, all lines remain to understand how synthetic data is used appropriately and improve user literacy toward synthetic data usage within the organisation.

In this sense, synthetic data governance becomes an application of the same collaborative model proposed in this study—not an addition to it.

In conclusion, the friction between transparency and confidentiality is not an inherent limitation of GRC, but rather a consequence of how data access is designed. This study demonstrates that synthetic data, as a controlled intermediary layer, enables the collaborative intent of the Three Lines Model to become operationally feasible. However, this solution carries its own governance requirements and inherent risks: the synthetic data lifecycle — from source data integrity, model development, and ongoing validation to user literacy across governance functions — must itself be governed with equal rigour. Crucially, this does not require a separate framework — the 3LM itself provides the accountability structure to govern it, with each line assuming responsibility for a distinct aspect of the synthetic data lifecycle. In this sense, synthetic data does not complicate governance; it deepens it.

Reproducibility

The full simulation pipeline — including synthetic data generation, governance simulation logic, and evaluation scripts — is made publicly available to support replication and extension of this study. The codebase can be accessed at: https://github.com/dphir/thesis-sdg.

All experiments apply a fixed random seed of 42 throughout sampling, model training, and generation to ensure full reproducibility. The software environment is detailed in Section 4.

The original dataset cannot be shared due to privacy constraints. Furthermore, the simulation logic — including threshold parameters, baseline calibration, and escalation rules — was designed based on the characteristics of the available dataset and use case. Researchers applying this pipeline to different datasets or operational contexts should expect to adjust these parameters accordingly.

Acknowledgements

We wish to express sincere gratitude to the supervisors, Mahmoud Shoush, PhD, and Eduardo Ribas Brito, MSc, for their critical feedback, constructive guidance, and thoughtful insights throughout the research and writing process.

Special thanks are extended to Fredrik Milani, whose willingness to provide access to the dataset that made this study possible is deeply appreciated.

We also wishes to thank the staff of the University of Tartu for their consistent support throughout the programme, as well as my mentor from the University of Tartu mentorship programme for their encouragement and guidance.

Finally, heartfelt gratitude goes to our family and colleagues, whose unwavering support and encouragement made the completion of this thesis possible.

References

[1] Institute of Internal Auditors. The IIA's Three Lines Model: An update of the Three Lines of Defense. Institute of Internal Auditors, July 2020. https://www.theiia.org/en/content/position-papers/2020/the-iias-three-lines-model-an-update-of-the-three-lines-of-defense/.

[2] Bantleon U., d'Arcy A., Eulerich M., Hucke A., Pedell B., and Ratzinger-Sakel N. V. S. Coordination challenges in implementing the three lines of defense model. *International Journal of Auditing* 25.1 (2021), pp. 59–74. DOI: 10.1111/ijau.12201.

[3] Sillaber C., Boticiu S., Idelberger F., and Stocker C. Experience: Data and information quality challenges in governance, risk, and compliance management. *Journal of Data and Information Quality* 11.2 (2019). DOI: 10.1145/3297721.

[4] Deloitte Middle East. The integrated assurance journey: A point of view on navigating the complexity of modern risk landscapes. Tech. rep. Point of View Edition 48. Deloitte, 2025. https://www.deloitte.com/content/dam/assets-zone2/middle-east/en/docs/generic/2025/me-pov-48/the-integrated-assurance-journey-pov48.pdf.

[5] Sahoo N. DORA, NIS2, and GDPR: The New Compliance Power Triangle Reshaping the EU. LinkedIn Resource: Cybersecurity and Compliance Frameworks. Feb. 2024. https://www.linkedin.com/pulse/dora-nis2-gdpr-new-compliance-power-triangle-reshaping-narendra-sahoo-0yucc/ (10/23/2025).

[6] European Data Protection Supervisor. TechSonar: Synthetic Data. Accessed: 2025. 2024. https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en.

[7] Financial Conduct Authority. Synthetic Data for Models in Financial Services: Governance Considerations. Tech. rep. FCA, 2025. https://www.fca.org.uk/publication/corporate/synthetic-data-models-financial-services-governance.pdf.

[8] Teichmann F., Boticiu S., and Sergi B. S. RegTech — Potential benefits and challenges for businesses. *Technology in Society* 72 (2023), p. 102150. DOI: 10.1016/j.techsoc.2022.102150.

[9] Rafiq M. and Sohail M. K. Adopting regulatory technology for anti-money laundering in banking: Key enablers and barriers in a RegTech model. *Sustainable Futures* (2025). DOI: 10.1016/j.sftr.2025.101377.

[10] Deloitte. Modernizing the three lines of defense model. Deloitte, 2021. https://www2.deloitte.com/us/en/pages/financial-services/articles/modernizing-the-three-lines-of-defense-model.html.

[11] Kaminski P. and Schonert J. The neglected art of risk detection. Tech. rep. Risk Practice. McKinsey & Company, Mar. 2017. https://www.mckinsey.de/~/media/McKinsey/Business%20Functions/Risk/Our%20Insights/The%20neglected%20art%20of%20risk%20detection/The-neglected-art-of-risk-detection.pdf.

[12] Pilgram L., Fineberg A., Jonker E., and El Emam K. An Assessment of Synthetic Data Generation, Use and Disclosure Under Canadian Privacy Regulations. *International Journal of Population Data Science* (2024). DOI: 10.23889/ijpds.v9i4.2342.

[13] PA Global. SOX: Restoring trust after corporate scandals. Insights on Sarbanes-Oxley and Corporate Governance. PA Global, June 2022. https://pa-global.com/insights/sox-restoring-trust-after-corporate-scandals/.

[14] OCEG. What is GRC (Governance, Risk, and Compliance)? OCEG, 2020. https://www.oceg.org/what-is-grc/.

[15] OCEG. GRC Capability Model 3.5 (Red Book). OCEG, 2021. https://www.oceg.org/grc-capability-model-red-book/.

[16] Committee of Sponsoring Organizations of the Treadway Commission. Internal control — Integrated framework. COSO, 2013. https://www.coso.org/.

[17] SAP. What is GRC (Governance, Risk, and Compliance)? Enterprise Resource Planning and Compliance Resources. SAP, 2024. https://www.sap.com/resources/what-is-grc.

[18] Financial Action Task Force. Risk-Based Approach for the Banking Sector. Tech. rep. FATF/OECD, 2014. https://www.fatf-gafi.org/content/dam/fatf-gafi/guidance/Risk-Based-Approach-Banking-Sector.pdf.

[19] Gadimov E. and Birihanu E. Real-time suspicious detection framework for financial data streams. *International Journal of Information Technology* (2025). DOI: 10.1007/s41870-025-02529-6.

[20] Dankar F. K. and Mahmoud I. Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. *Applied Sciences* 11.5 (2021). [VERIFICATION: CONFIRMED. This entry matches the citation provided directly in the source text: “K. Dankar, I. Mahmoud. Fake it till you make it: guidelines for effective synthetic data generation, Applied Sciences 11.5 (2021): 2158, 2021.” DOI verified via MDPI.], p. 2158. DOI: 10.3390/app11052158.

[21] Valkenburg B. and Bongiovanni I. Unravelling the three lines model in cybersecurity: a systematic literature review. *Computers & Security* 139 (2024), p. 103708. DOI: 10.1016/j.cose.2024.103708.

[22] Institute of Internal Auditors. The Three Lines of Defense in Effective Risk Management and Control. Position Paper. The Institute of Internal Auditors, Jan. 2013. https://theiia.fi/wp-content/uploads/2017/01/pp-the-three-lines-of-defense-in-effective-risk-management-and-control.pdf.

[23] Reite E. J., Oust A., Bang R. M., and Maurstad S. Changes in credit score, transaction volume, customer characteristics, and the probability of detecting suspicious transactions. *Journal of Money Laundering Control* 26.6 (2023), pp. 1165–1178. DOI: 10.1108/JMLC-06-2022-0087.

[24] Raghunathan T. E. Synthetic Data. *Annual Review of Statistics and Its Application* 8 (2021). [VERIFICATION: CONFIRMED. This entry matches the citation provided directly in the source text: "T. E. Raghunathan, Synthetic data, Annual Review of Statistics and Its Application, 8, 129-140, 2021."], pp. 129–140. DOI: 10.1146/annurev-statistics-040720-031516.

[25] Hradec J., Craglia M., Di Leo M., De Nigris S., Ostlaender N., and Nicholson N. Multipurpose Synthetic Population for Policy Applications. EUR 31116 EN. [VERIFICATION: CONFIRMED. Full citation including ISBN and JRC number (JRC128595) is provided directly in the source text. DOI verified.] Luxembourg: Publications Office of the European Union, 2022. DOI: 10.2760/50072.

[26] Hittmeir M., Ekelhart A., and Mayer R. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. *Applied Sciences* 11.5 (2019), p. 2158. DOI: 10.3390/app11052158.

[27] Assefa S. A., Dervovic D., Mahfouz M., Tillman R. E., Reddy P., and Veloso M. Generating Realistic Stock Market Order Streams. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. [VERIFICATION: PLAUSIBLE. The source text cites "Assefa et al." in the context of financial synthetic data (TimeGAN/diffusion). The entry above is a well-known Assefa et al. paper on synthetic financial data; however, the source text does not provide the full citation. Verify this is the intended reference.] 2020, pp. 1–8. DOI: 10.1609/aaai.v34i01.5415.

[28] Kotelnikov A. et al. TabDDPM: Modelling Tabular Data with Diffusion Models. *arXiv preprint arXiv:2209.15421* (2023). https://arxiv.org/abs/2209.15421.

[29] Chen T.-Y., Liu C.-Y., Li W., and Chan Y.-L. Balancing Class Imbalance in Anti-Money Laundering with GAN-Based Synthetic Data. *IEEE Access* (2021). [VERIFICATION:

UNVERIFIED. The source text cites “Chen et al., 2021” for VAE- and GAN-based synthetic data for AML minority-class recall. The entry above is a representative placeholder. Locate the precise Chen et al. (2021) paper before submission.]

[30] Cardoso J., Saleiro P., and Bizarro P. LaundroGraph: Self-Supervised Graph Representation Learning for Anti-Money Laundering. *arXiv preprint* (2022). [VERIFICATION: PARTIALLY CONFIRMED. “Cardoso et al., 2022” is cited for graph-based AML. LaundroGraph (arXiv:2210.04285) is a prominent 2022 graph-learning AML paper. Confirm this is the intended reference; if a journal version exists, update accordingly.] arXiv: 2210.04285.

[31] Qian Z. et al. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. *arXiv preprint arXiv:2301.07573* (2023).

[32] Lautrup A. D., Hyrup T., Zimek A., and Schneider-Kamp P. SynthEval: A Framework for Detailed Utility and Privacy Evaluation of Tabular Synthetic Data. *Data Mining and Knowledge Discovery* (2024). DOI: 10.1007/s10618-024-01081-4.

section*Appendices

List of Acronyms

3LM Three Lines Model. 2–4, 10–15, 20, 21, 26, 27, 35, 36, 41, 43–48, 51–56, 59, 60

3LoD Three Lines of Defense. 2, 4, 10, 13, 15, 19, 20, 26, 27, 35, 36, 40, 43–49, 51, 52

AI Artificial Intelligence. 15, 22, 23

AML Anti-Money Laundering. 10, 11, 13, 16, 17, 22–24, 28, 29, 54, 58

COBIT Control Objectives for Information and Related Technologies. 40

DORA Digital Operational Resilience Act. 10, 13, 16, 55

DPIA Data Protection Impact Assessment. 55

DSS02 Deliver, Service, and Support-Managed Service Requests and Incidents. 41

EBA European Banking Authority. 55

FATF Financial Action Task Force. 29

FCA Financial Conduct Authority. 12

FRR Field Reduction Ratio. 36

GAN Generative Adversarial Network. 23, 24

GCR Granularity Compression Ratio. 34, 36, 47, 54

GDPR General Data Protection Regulation. 2, 4, 10, 13, 16, 18–20, 35, 52, 55, 56

GRC Governance, Risk, and Compliance. 2–4, 9–13, 15, 16, 18, 22, 23, 26, 49, 51–58, 60

IIA Institute of Internal Auditors. 10, 16, 20, 21, 35

ITIL Information Technology Infrastructure Library. 41

JSD Jensen-Shannon Divergence. 38–40, 46, 48, 50, 52

KS Kolmogorov-Smirnov. 38–40, 46, 48, 50, 52

KYC Know Your Customer. 11, 17, 58

ML Machine Learning. 48, 50, 54

NIS2 Network and Information Security Directive 2. 16

OCEG Open Compliance and Ethics Group. 16

PET Privacy-Enhancing Technology. 21

PII Personally Identifiable Information. 21, 35, 36, 40–48, 51

RBAC Role-Based Access Control. 54, 55

RDR Reconstruction Dependency Ratio. 34, 37, 41, 42, 47, 54

RegTech Regulatory Technology. 11, 17

SDG Synthetic Data Generation. 9, 11, 12, 15, 18, 21–24, 31, 36, 40, 42, 46, 47, 54, 55

SLA Service Level Agreement. 40

SOP Standard Operational Procedure. 48

TabDDPM Tabular Denoising Diffusion Probabilistic Model. 31, 33

VAE Variational Autoencoder. 23, 24

Table 14. AI-assisted activities outside the experimental pipeline. All outputs were reviewed and, where necessary, rewritten by the author before inclusion.

Activity	**Tool(s)**	**Scope & Author Review**
Literature synthesis & summarisation	NotebookLM	Used for organising related papers, mind-mapping structures, and identifying research gaps. All highlighted points were verified against primary sources by the author.
Writing assistance (copy-editing, translation & stylistic refinement)	Claude, Gemini, Grammarly, overleaf AI	Corrected grammar and typography; translated author-written ideas into academic English. No new claims were introduced; all revisions reviewed and adjusted manually by the author.
Diagram & visual code generation	Claude, Gemini	Converted author-sketched workflows into TikZ/Mermaid/HTML visualisation code. All diagram structures, labels, and layouts were manually refined by the author prior to inclusion.
Simulation design & code development	Claude	The author specified all simulation logic, threshold parameters, and pipeline structure. AI assisted in translating specifications into boilerplate code and served as a sounding board for methodology stress-testing. All code was reviewed and tested against expected outputs using fixed random seeds.
Results tabulation & governance metric framework	Claude	AI formatted simulation outputs into summary tables and helped structure evaluation dimensions into measurable indicators. All values, metric definitions, and analytical interpretations were independently determined and verified by the author.

License

Non-exclusive licence to reproduce the thesis and make the thesis public

I, **Defir Ilmi Faridha**,

1. grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis

   From Data Silos to Integrated Assurance: A Synthetic Data Approach to the Three Lines Model for Governance, Risk, and Compliance,

   supervised by **Mahmoud Shoush, PhD ; Eduardo Ribas Brito, MSc**;

2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;
3. am aware of the fact that the author retains the rights specified in points 1 and 2;
4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Defir Ilmi Faridha

14/05/2026