EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Alessandra Agati, Adele Fontana,
João Matos Leite, Claudia Schwarz

Unlocking supervisory effectiveness in credit risk: insights from the ECB's supervisory findings and measures

No 3308 / November 2026



Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Abstract

This paper provides an assessment of supervisory effectiveness in credit risk using confidential data from ECB Banking Supervision. The analysis covers 8,106 supervisory measures issued and closed between 2016 and 2025. A central contribution is the construction of a measure-level indicator of supervisory effectiveness based on large language model (LLM) analysis of supervisory narratives and closure justifications. The indicator captures whether supervisory measures resulted in full remediation of the underlying deficiencies. Using logistic and duration models, we examine the characteristics of findings, measures, and banks associated with remediation outcomes and remediation speed. The results show that multi-faceted findings are a key driver of supervisory outcomes. Measures related to data, reporting, and IT deficiencies are less likely to achieve full remediation and require longer remediation periods. Measures originating from on-site inspections are associated with higher remediation rates and faster closure. Stronger governance is also associated with more successful remediation outcomes.

Keywords: supervisory effectiveness, credit risk, large language models, banking supervision

JEL classifications: G21, E51, H81, G28

Non-technical summary

Effective banking supervision requires more than identifying weaknesses in banks. It also requires ensuring that banks address those weaknesses in a timely and comprehensive manner. Yet assessing whether supervisory actions actually lead to remediation remains challenging. Supervisory performance is often measured through completed processes and activities, while evidence on whether supervisory interventions resolve the underlying deficiencies is much more limited.

This paper provides new evidence on supervisory effectiveness in the area of credit risk using confidential data from ECB Banking Supervision. The analysis is based on more than 8,100 supervisory measures issued and closed between 2016 and 2025 and recorded in the ECB's Findings and Measures database. The paper focuses on whether supervisory measures achieved their intended objective, namely the remediation of the deficiencies identified by supervisors.

A key challenge is that remediation outcomes are not directly recorded in structured supervisory data. Although a measure may be formally closed, closure does not necessarily imply that the underlying issue has been fully resolved. To address this challenge, the paper applies large language models (LLMs) to supervisory narratives and closure justifications to classify measures according to whether they resulted in full, partial, or no remediation. This creates a novel outcome-based indicator of supervisory effectiveness that can be analysed systematically across a large number of supervisory cases.

Using this indicator, the paper examines which characteristics of supervisory measures, findings, and banks are associated with successful remediation and the time required to achieve it.

The results show that supervisory effectiveness differs substantially across both measure types and institutions.

First, the complexity of the underlying finding is an important determinant of outcomes. Measures addressing multiple issues are less likely to achieve full remediation and take longer to close. Similarly, deficiencies related to data, reporting and IT systems are among the most difficult to remediate, both in terms of success rates and remediation time. This reflects the substantial organisational, technological, and operational changes often required to address such weaknesses.

Second, measures originating from on-site inspections are associated with more successful and faster remediation than measures resulting from broader supervisory processes such as the Supervisory Review and Evaluation Process (SREP) or ongoing supervisory monitoring. This finding suggests that highly targeted and in-depth supervisory assessments can facilitate corrective action and highlights the importance of using different supervisory tools in a complementary manner.

Third, more severe supervisory measures are less likely to achieve full remediation and generally remain open for longer periods. This likely reflects the fact that more severe findings are often associated with deeper-rooted weaknesses and require more extensive corrective action.

Fourth, bank characteristics matter, although their influence is generally weaker than that of finding and measure characteristics. Banks with stronger governance frameworks and those with higher profitability tend to achieve higher remediation rates. However, some of these relationships become less pronounced once differences across supervisory teams and institutions are taken into account, suggesting that remediation outcomes are influenced by a combination of bank-specific and supervisory factors.

The results have several implications for supervisory authorities. They highlight the importance of designing supervisory measures that are sufficiently focused, clearly communicated, and tailored to the nature of the underlying deficiency. They also point to the particular challenges posed by complex, cross-cutting issues, especially those involving data and IT infrastructures, where remediation may require longer time horizons and sustained supervisory follow-up. More generally, the results can support supervisors in allocating resources and calibrating supervisory intensity more effectively.

Beyond the substantive findings, the paper demonstrates how artificial intelligence can support supervisory effectiveness assessments. By transforming large volumes of unstructured supervisory information into structured data, LLMs make it possible to evaluate supervisory outcomes at a level of granularity that would otherwise be difficult to achieve. While the results should be interpreted as evidence of associations rather than causal effects, they provide a new and practical framework for monitoring and analysing supervisory effectiveness in banking supervision.

1. **Introduction**

The Global Financial Crisis and subsequent episodes of banking stress highlighted the importance of effective banking supervision and triggered far-reaching reforms of regulatory and supervisory frameworks. These reforms included revisions to the Capital Requirements Directive and Capital Requirements Regulation, as well as the establishment of the Single Supervisory Mechanism (SSM) in 2014 under the auspices of the European Central Bank (ECB). More recently, the banking turmoil of March 2023 renewed attention on the role of supervision in identifying vulnerabilities and ensuring that banks address them in a timely manner.

But what defines effective supervision, and how can it be assessed? The Basel Committee on Banking Supervision (BCBS, 2025) defines supervisory effectiveness as: *"Effective supervision promotes the safety and soundness of banks and the banking system by promptly assessing prudential risks, identifying material shortcomings within banks, and using the supervisory toolkit and powers appropriately to ensure that banks remediate shortcomings in a timely manner."* Effectiveness extends beyond compliance, focusing on timely intervention, corrective action, and institutional resilience. The BCBS identifies two key purposes for assessing supervision: public accountability and optimising supervisory strategy, such as determining which tools to use for different banks and how to allocate resources. Crucially, effectiveness should be measured by outcomes, not just outputs, as what matters is whether a bank's remedial actions resolve identified shortcomings.

Importantly, supervisory effectiveness should not be equated with changes in observed risk metrics. Supervisory measures rarely seek to reduce a specific level of credit risk directly. Instead, they typically aim to address weaknesses in governance arrangements, risk-management processes, internal controls, methodologies, data quality, reporting frameworks, or other prudential deficiencies. Consequently, a supervisory measure should be considered effective when the bank successfully implements the required corrective actions and remediates the identified shortcoming. In fact, effective supervision may sometimes lead to a temporary increase in measured risk if improved methodologies, controls, or reporting practices result in more accurate risk identification. For this reason, the focus of this paper is on remediation outcomes rather than subsequent changes in credit risk indicators or prudential metrics.

Despite a growing literature on supervisory effectiveness, evidence on supervisory outcomes remains limited. Existing studies typically evaluate supervisory tools by examining their impact on bank behaviour, lending activity, risk-taking, or market outcomes. While these approaches provide valuable insights, they do not address a more granular question: whether individual supervisory measures achieve the remediation of the specific deficiencies they were intended to correct. Answering this question has proved difficult because the relevant information is rarely available in structured datasets and often remains embedded in confidential supervisory documentation.

This paper addresses that gap using confidential data from ECB Banking Supervision's Findings and Measures (F&M) database. The database contains detailed information on supervisory findings and measures generated through on-site inspections, horizontal reviews, ongoing supervisory activities, and the Supervisory Review and Evaluation Process (SREP). Although the database contains more than 75,000 observations since the establishment of the SSM, much of the relevant information is stored in unstructured narrative fields, making systematic analysis challenging.

To overcome this challenge, we apply large language models (LLMs) to transform supervisory narratives into structured analytical variables. We first classify supervisory findings and measures into thematic categories using a dedicated taxonomy framework. We then construct a novel measure-level indicator of supervisory effectiveness by analysing supervisory closure narratives and classifying measures according to whether they resulted in full, partial, or no remediation of the underlying deficiency. This approach allows us to move beyond traditional process-based measures of supervision and towards an outcome-oriented assessment of supervisory effectiveness.

Our analysis focuses on 8,106 credit-risk-related supervisory measures issued and closed between 2016 and 2025. Using the LLM-based effectiveness indicator, we examine which characteristics of supervisory measures, findings, and supervised institutions are associated with successful remediation in a logistic regression framework. We then analyse the duration of fully effective measures to identify the factors associated with faster or slower remediation using Cox proportional hazards model.

Three main findings emerge. First, finding complexity is an important determinant of supervisory outcomes. Measures addressing multiple issues are less likely to achieve full remediation and remain open for longer periods. Data, reporting, and IT-related deficiencies

also prove particularly difficult to remediate. Second, supervisory measures arising from on-site inspections exhibit higher remediation rates and shorter remediation periods than measures originating from broader supervisory processes such as the SREP or ongoing supervisory monitoring. Third, bank characteristics matter, albeit to a lesser extent than finding and measure characteristics. Stronger governance and greater profitability are associated with better remediation outcomes, although these relationships weaken once supervisory-team heterogeneity is considered.

This paper contributes to the literature in three ways. First, it provides one of the first large-scale empirical assessments of supervisory effectiveness at the level of individual supervisory measures rather than supervisory frameworks or instruments. Second, it introduces a novel outcome-based measure of supervisory effectiveness grounded in the remediation of identified deficiencies. Third, it demonstrates how artificial intelligence can be used to extract policy-relevant information from large volumes of confidential supervisory text, thereby expanding the scope for empirical research and policy evaluation in banking supervision.

The remainder of the paper is organised as follows. Section 2 reviews the literature on supervisory effectiveness and textual analysis. Section 3 describes the methodology and the construction of the LLM-based effectiveness indicator. Section 4 presents the data. Section 5 discusses the empirical results. Section 6 concludes.

2. Literature review

This study draws on two strands of research. The first is the empirical literature on the effectiveness of banking supervision, reviewed in Section 2.1. The second is the literature on textual data and artificial intelligence in economics and finance, reviewed in Section 2.2. This paper is at the intersection of the two: a granular, measure-level assessment of supervisory effectiveness, made possible by applying large language models to a large body of unstructured supervisory text.

2.1 Evaluating supervisory effectiveness

The most comprehensive recent survey of this field is the Basel Committee's literature review on supervisory effectiveness (BCBS, 2025). It organises the evidence around a 'house of effectiveness': a set of enablers and impediments at the base, three pillars: risk identification and assessment, remediation and enforcement, and collaboration and transparency and

supervisory culture as the roof. This section follows that structure, after first setting out what effectiveness means and why it is hard to measure.

2.1.1 Risk identification and assessment

The first pillar concerns the tools through which supervisors identify and assess risk: on-site inspections, supervisory stress tests, horizontal reviews and benchmarking, and early-warning models (BCBS, 2025). On-site inspections have received the most empirical attention, because they are intrusive and may involve a detailed review of the loan book. The Basel Committee's review reports that inspections promote more conservative risk management, with more prudent provisioning afterwards, and that the supervisory ratings derived from them have predictive power for financial stress (BCBS, 2025). Consistent with this, Passalacqua et al. (2021) use the quasi-random selection of Italian banks for inspection and find that inspected banks are more likely to reclassify loans as non-performing, cut credit to unproductive 'zombie' firms, and reallocate lending towards more productive borrowers, with positive spillovers for the wider economy; they trace these effects to improvements in bank governance. Bonfim et al. (2023) report comparable results for inspections of the largest Portuguese banks, although in that case the effect was confined to the inspected portfolios. Importantly, supervisory activities often extend beyond risk identification and contribute directly to remediation. The distinction between assessing weaknesses and ensuring their correction is therefore not always clear-cut.

2.1.2 Remediation and enforcement

The second pillar concerns the tools through which supervisors ensure that banks correct identified weaknesses and comply with prudential requirements. These include capital constraints under Pillar 2, enforcement and sanction actions, early-intervention regimes, and individual accountability regimes (BCBS, 2025). For a study focused on credit risk, the evidence on sanctions is particularly relevant. The Basel Committee's working paper finds that enforcement actions promote a more truthful classification of credit quality, reduce bank risk and lending, and generate deterrent effects on non-sanctioned banks, with effectiveness increasing in severity and depending critically on timing (BCBS, 2025).

Prior studies find that supervisory sanctions and capital-related supervisory measures can reduce risk-taking and strengthen resilience, although their effectiveness depends on timing, severity and institutional context (Delis et al., 2016; Caiazza et al., 2018; Berger et al., 2022; BCBS, 2025). This literature is the closest precedent for the present study, yet a gap remains.

Each of these studies measures the response of the bank, or of credit supply, to an enforcement tool. None measures whether a specific supervisory measure was implemented, and the underlying shortcoming rectified. Whether a measure achieves its purpose is unlikely to be uniform: a narrowly targeted measure may be more readily remedied than a broad one, and a bank with stronger governance, may comply more readily than a weaker peer.

These remain conjectures because the evidence does not reach the level of the individual measure. As a result, little is known about which categories of findings are remediated, which are not, and which characteristics of banks and supervisory measures distinguish the two outcomes.

Other dimensions of supervisory effectiveness, including coordination, transparency, consistency and supervisory culture, have also received attention in the literature. For example, Agarwal et al. (2014) show that differences in supervisory practices across authorities can lead to inconsistent regulatory outcomes, while the BCBS (2025) highlights the roles of supervisory cooperation, transparency and organisational culture in supporting effective oversight. These dimensions are important determinants of supervisory outcomes but are less directly related to the measure-level focus of this paper.

Taken together, the literature shows that the effectiveness of supervision depends on the tools used, on the bank, and on the institutional setting. What it does not provide is evidence at the level of the individual measure and its remediation. The binding constraint is not a lack of observations, but their form and confidentiality: the relevant information is recorded as free text rather than structured data and is typically not publicly accessible. This limitation is therefore less about data availability than about the ability to analyse them systematically. Turning unstructured supervisory text into analysable variables, and so into outcome-level evidence, is the methodological problem addressed in the next section.

2.2 Textual analysis and artificial intelligence

A key reason why effectiveness remains difficult to assess is that some of the richest supervisory records are stored as free text, inspection reports, findings, and the narratives that justify supervisory decisions, which conventional statistical methods cannot readily exploit. Advances in textual analysis, particularly through artificial intelligence, make it possible to convert this information into structured data, thereby enabling large-scale empirical analysis of supervisory outcomes that would otherwise be prohibitively time-consuming. Text-as-data approaches have become increasingly common in economics and finance, enabling

researchers to extract information from large collections of documents that would otherwise be difficult to analyse systematically (Gentzkow et al., 2019).

Early applications relied primarily on dictionary-based approaches, which classify text using predefined word lists. While useful, these methods often struggle to capture context-specific meanings in financial and supervisory language (Loughran and McDonald, 2011). More recently, machine learning techniques and large language models (LLMs) have substantially expanded the scope of textual analysis by allowing information to be interpreted in context and classified with greater flexibility (Huang et al., 2023; BIS, 2024; Korinek, 2023).

These methods are increasingly being applied to central banking and financial stability analysis. Studies have shown that textual information contained in central bank publications can provide useful signals regarding perceptions of risk and future financial vulnerabilities (Correa et al., 2021). Recent work at the ECB further demonstrates the potential of LLMs to identify and quantify risk assessments in financial stability communication and related policy documents (Fell et al., 2026). Applications to supervisory records remain comparatively limited, largely because such information is confidential. One notable exception is Goldsmith-Pinkham, Hirtle and Lucca (2016), who analyse the content of US supervisory examination reports and show that supervisory narratives contain information beyond formal supervisory ratings. However, existing studies have generally focused on characterising supervisory communication rather than assessing the outcomes of supervisory actions.

2.3 Research gap and contribution

The two strands reviewed above leave a specific gap. While the supervisory effectiveness literature provides evidence on the impact of broad supervisory tools, it rarely assesses whether individual measures lead to the remediation of underlying deficiencies. This reflects the distinction between supervisory outputs, i.e. completed supervisory actions, and supervisory outcomes, i.e. effective remediation. At the same time, textual-analysis methods have rarely been applied to the outcomes of confidential supervisory actions.

Using confidential supervisory data from ECB Banking Supervision, we apply large language models to identify whether individual measures resulted in full, partial or no remediation. This allows an outcome-oriented assessment of supervisory effectiveness at the measure level and provides evidence on the characteristics of supervisory measures and banks associated with more effective remediation outcomes.

3. Methodology

This section describes the methodology used to extract supervisory information from textual records through large language models and to quantify supervisory effectiveness at the measure level.

3.1 Constructing the Effectiveness Variable Using Large Language Models

A key empirical challenge is that supervisory effectiveness, i.e. whether a measure achieves its intended objective, is not directly observable in structured supervisory data. While the F&M database records supervisory actions and their administrative status, the closure of a measure does not necessarily imply that the underlying deficiency has been fully remediated. Measures may be closed despite only partial remediation, because they are superseded by new measures, or for other supervisory reasons. Consequently, a measure's closure status cannot be interpreted as a direct indicator of supervisory effectiveness.

The information required to assess remediation is instead contained in qualitative supervisory narratives, particularly in the descriptions of identified shortcomings and in the justifications provided when measures are closed. The objective of the LLM-based framework is therefore to transform this information into a systematic measure of whether the supervisory objective was ultimately achieved.

The construction of the outcome variable is grounded in the definition of supervisory effectiveness articulated by the BCBS. Importantly, the outcome variable is not designed to measure changes in a bank's risk profile as this is typically not the intended objective of the supervisor. Rather, the effectiveness measure developed in this paper captures whether the supervisory objective underlying a measure was achieved, namely whether the identified deficiency was adequately remediated by the supervised entity. The analysis therefore does not seek to estimate the impact of supervisory measures on realised credit-risk outcomes or broader prudential indicators. Instead, it evaluates supervisory effectiveness at the level of the individual supervisory intervention, focusing on whether the action achieved the corrective outcome sought by supervisors.

In line with the focus of this paper, we adopt a narrower operational definition: a measure is considered effective if it leads to the full remediation of the underlying deficiency identified by the supervisor. Measures that are only partially implemented, or for which remediation remains

incomplete at closure, are classified as *somewhat effective*, while measures for which no remediation is observed are classified as *ineffective*.

We implement the classification task using the GPT-4.1 model in a prompt-based framework. Each observation consists of two textual inputs: (i) the measure description, outlining the identified shortcoming and supervisory expectations, and (ii) the closure narrative, describing the actions taken by the bank and the rationale for closing the measure. Based on this information, the model determines whether the deficiency has been effectively resolved. The temperature parameter is set to 0.15 to promote stable and consistent outputs.

We encode the operational definition of effectiveness in the prompt and instruct the model to base its assessment exclusively on the information contained in the input texts. We restrict the model output to three categories, *effective*, *somewhat effective*, and *ineffective*, ensuring consistency across observations. This approach embeds expert judgement ex ante and applies it uniformly across the dataset.

Given the model-generated nature of the dependent variable, several validation steps are performed. First, we conduct human-in-the-loop validation on selected subsamples and use expert assessments to refine the prompt. Second, to reduce stochastic variation, we repeat the classification ten times and aggregate the results using a majority-voting rule. Third, we replicate the exercise using an alternative model (o4-mini) yielding an agreement rate of approximately 90%.

Despite these safeguards, the approach remains subject to limitations. The classification depends on the completeness and clarity of supervisory narratives and may therefore be affected by reporting practices. In addition, prompt-based methods are inherently sensitive to instruction design and model-specific behaviour. Finally, the resulting indicator captures documented remediation rather than the broader impact of supervisory measures on bank risk.

Moreover, supervisory findings and associated measures are classified into six remediation-topic categories using a dedicated LLM taxonomy-classification framework: Governance & Controls; Classification & Staging; Data, Reporting & IT; Provisioning & Model Risk; NPE & Collateral Management; and Credit Lifecycle Management. The taxonomy-development process, category definitions, and validation procedures are reported in Technical Annex B.

3.2 Empirical Strategy

3.2.1 Drivers of effectiveness: binomial logit

Having constructed a measure-level indicator of supervisory effectiveness using an LLM-based classification approach, we turn to the empirical analysis of its determinants. The objective is to assess which characteristics of supervisory measures and supervised institutions are associated with full remediation of the underlying issues.

In the main specification the dependent variable, *Effectiveness*, is defined as a binary indicator equal to one if a supervisory measure is classified as *effective* (full remediation of the underlying supervisory concern) and zero if it is classified as *somewhat effective* or *ineffective*. This specification focuses on the factors associated with achieving complete remediation, which represents the supervisory outcome of primary interest.

The binary specification is preferred as the baseline model because the distinction between full remediation and all other outcomes directly aligns with the supervisory objective of achieving complete resolution of identified deficiencies, while also reducing concerns related to the relatively smaller number of observations in the ineffective category.

To analyse the determinants of effectiveness, we estimate a binary logistic regression model of the following form:

$$\Pr\left(Effectiveness_{m,b,t} = 1\right) = \Lambda(\alpha + Finding\ \&\ Measure\ characteristics_{m,b,t}\beta + Bank\ characteristics_{b,t-1}\gamma + \delta_t)$$

where *m*, *b*, and *t* index the supervisory measure, the bank, and time (quarter), respectively. The vector Finding and M*easure characteristics* includes characteristics of the supervisory action, such as its legal enforceability, source, severity, and the type and number of findings identified through the LLM-based classification framework. The vector *bank characteristics* captures bank-level characteristics measured one quarter before the time of issuance of the supervisory measure, including indicators of credit risk materiality, profitability, and governance. All specifications include year fixed effects to control for common macro-financial conditions and changes in supervisory practices over time.

As with any observational analysis, several limitations should be acknowledged. First, unobserved heterogeneity may affect the estimated relationships, particularly through

differences in supervisory intensity and supervisory practices across Joint Supervisory Teams (JSTs). Second, the estimation sample is not randomly drawn. Supervisory findings are more likely to be identified in institutions that are subject to closer supervisory scrutiny, including through on-site inspections and horizontal reviews. Third, the effectiveness classification relies on the information contained in supervisory documentation and on the ability of the LLM-based framework to consistently interpret those texts.

To assess the robustness of the results, we perform several additional analyses. First, we introduce JST and country fixed effects. Second, we re-estimate the model excluding ineffective measures. Then, we estimate a multinomial logit specification distinguishing between *ineffective*, *somewhat effective*, and *effective* outcomes. This alternative specification allows the determinants of full, partial, and absent remediation to differ and provides a robustness check for the simplifying assumption embedded in the binary classification.

3.2.2 Drivers of measure duration: Cox proportional hazards model

To explore the relationship between measure and bank characteristics and the time an effective measure remains open, we rely on a proportional hazard Cox model, of the following form:

Equation 1: $Cox\ proportional\ hazards\ model$

$$\lambda(Measure\ Duration|C_{m,b}) = \lambda_0(Measure\ Duration) \cdot exp(\beta_1 C_{m,b})$$

Where $Measure\ Duration$ is the number of days a measure remains open before remediation, $C_{m,b}$ are the measure and bank characteristics and $\lambda_0(Measure\ Duration)$ is the baseline hazard model.

In other words, we aim at understanding to which extent each measure and bank characteristic at the opening of the measure influences the remediation time.[2]

Since the baseline hazard model is unknown, to estimate the full hazard function in equation 2 we first estimate β_1 via a Cox model, and, assuming no time-dependency on the right-hand side, take the hazard ratios between measures. The resulting hazards are proportional across measures. The measure hazard ratio, $\exp(\beta_1 C_{m,b})$, is however not telling on its own, to interpret the results, we rely instead on $\exp(\beta_1)$ and on the ratios of hazards between observed

[2] Since we are exploring the drivers of duration on the subset of fully effective measures (as categorised by LLM), the hazard model does not include censored measures, meaning open measures.

measures. This specification allows a flexible, semi-parametric estimation of relative hazard rates without imposing strong distributional assumptions on the baseline hazard.

However, in this setting the absence of censored observations limits the extent to which the model can exploit its core advantage of comparing measures that reach completion with those still ongoing over time.

In addition, as the sample is restricted to measures that achieve full remediation, the estimated effects should be interpreted as conditional on successful completion and are not representative of the overall remediation process.

To address these limitations and ensure robustness, we complement the main analysis with an Accelerated Failure Time (AFT) specification, as reported in equation 3, which directly models the duration of remediation and provides coefficients that can be interpreted in terms of proportional changes in completion time.

Equation 2: Accelerated Failure Time

$$log(Measure\ Duration_{m,b}) = C_{m,b}{}'\beta + \varepsilon_{m,b}$$

Notably, unlike the Cox model, in AFT a positive (negative) coefficient implies longer (shorter) duration.

More precisely, the acceleration factor is: $AF = e^{\beta}$

- $AF > 1$ duration increases,
- $AF < 1$ duration decreases.

3.3 Limitations

The empirical analysis is observational and is not designed to identify causal effects of supervisory measures on remediation outcomes. Accordingly, all results should be interpreted as conditional associations between supervisory measure characteristics, bank characteristics, and remediation outcomes, rather than as evidence of causal effect on supervisory effectiveness.

Several potential sources of endogeneity remain. First, supervisory measures are not assigned randomly. The type, severity, legal enforceability, and timing of supervisory actions are determined by supervisors in response to institution-specific circumstances. These same underlying circumstances may also influence the probability and speed of remediation. As a

result, the estimated coefficients may capture both the effect of the supervisory measure itself and the effect of the underlying conditions that prompted supervisory intervention.

Second, unobserved bank characteristics may influence both supervisory assessments and remediation capacity. To the extent that these factors are not fully captured by the available control variables, omitted variable bias may remain.

Third, supervisory intensity may vary across Joint Supervisory Teams (JSTs), reflecting differences in supervisory judgement, experience, resources, monitoring practices, or risk tolerance. To assess the relevance of such heterogeneity, robustness checks were performed by augmenting the baseline specification with JST-level and country-level fixed effects. The main findings remain qualitatively unchanged, particularly with respect to the effects of topic category, measure severity, supervisory originator, and measure vintage. This suggests that the core results are not primarily driven by persistent differences across supervisory teams or jurisdictions. Nevertheless, fixed effects cannot fully account for time-varying supervisory practices or other unobserved factors that may influence both supervisory actions and remediation outcomes.

At the same time, some features of the supervisory process reduce concerns regarding reverse causality. Supervisory findings originate from independent supervisory assessments rather than from bank self-reporting and are therefore, to some extent, exogenous to banks' subsequent remediation decisions. Nevertheless, residual confounding cannot be excluded. For this reason, the results are interpreted as descriptive evidence on the correlates of supervisory effectiveness rather than estimates of causal supervisory effects.

4. Data

The empirical analysis is based on supervisory data from the F&M database, complemented with bank-level information from supervisory reporting data.

The sample is restricted to measures related to credit risk, excluding internal model investigations findings,[3] with a final status of “closed” or “cancelled”. To ensure sufficient data quality and coverage, observations from 2014 and 2015 are excluded from the analysis. In addition, measures issued so far in 2026 are not part of this analysis, as many remain ongoing

[3] IMI findings are excluded because they are subject to a distinct supervisory framework, with specific assessment methodologies and remediation processes. Including them could bias the analysis of supervisory effectiveness by introducing observations that are not directly comparable to the broader population of credit risk measures.

and their full remediation cycle cannot yet be observed. This restriction ensures that the full remediation cycle can be assessed based on observable supervisory information.

The sample includes supervisory measures across all severity levels, allowing for a comprehensive assessment of supervisory effectiveness across the full spectrum of supervisory actions. The sample is further cleaned by excluding measures closed for procedural or administrative reasons and observations with insufficient information to assess remediation effectiveness. This ensures that the analysis is based only on measures for which remediation outcomes can be meaningfully evaluated.

After applying these filters, the final sample comprises 8,106 observations.

It is important to note that the sample reflects a selected subset of measures based on the applied filtering criteria. As a result, the sample may not be fully representative of the broader population of supervisory actions.

From the F&M database, we use several structured variables at the measure level. These include:

- the status of the measure (closed or cancelled),
- severity, distinguishing between low (M1), moderate (M2), high (M3) and very high (M4),
- legal enforceability, capturing whether the measure is a recommendation or an obligation. Recommendations are non-binding measures issued as part of the ongoing supervisory dialogue, whereas obligations are legally binding requirements imposed on the bank,
- the source of the finding, indicating the supervisory activity from which the issue originated. This includes:
    - on-site inspections, involving in-depth assessments conducted at the bank,
    - horizontal analyses, referring to cross-bank comparative reviews,
    - JST analyses, based on ongoing supervisory monitoring, and
    - SREP-related activities, which are part of the annual Supervisory Review and Evaluation Process and reflect a comprehensive assessment of a bank's risk profile and governance.

These variables capture key dimensions of supervisory action, including its origin, legal strength, and severity.

As an LLM-derived variable, we construct our main dependent variable, which captures the outcome of each supervisory measure in terms of remediation. Following the methodology described in Section 3.1, the variable is defined as a categorical indicator taking three values: effective, somewhat effective, and ineffective. A measure is classified as effective if the underlying supervisory deficiency is fully remediated at the time of closure; somewhat effective if remediation is only partial; and ineffective if no meaningful remediation is observed. The classification is based on the joint assessment of the measure description and the closure narrative using a prompt-based large language model.

In addition, we include as controls the type of finding, based on the main topic to be remediated, and the number of topics covered by a finding, both constructed using the LLM-based classification framework described in Section 3.1 and technical annex B. We use GPT-4.1 to classify findings into six categories covering the main dimensions of credit risk management.

The categories of findings we have identified are the following:

- Governance & Controls: Deficiencies related to credit risk governance, oversight, organisational arrangements, accountability, and internal control frameworks, including the effectiveness of management bodies and control functions.
- Classification & Staging: Weaknesses in the identification, classification, and regulatory treatment of credit exposures, including default recognition, NPE identification, forbearance assessment, and IFRS 9 staging.
- Data, Reporting & IT: Issues related to data quality, reporting capabilities, data governance, system integration, automation, and the IT infrastructure supporting credit risk management.
- Provisioning & Model Risk: Deficiencies in provisioning and impairment frameworks, expected credit loss methodologies, and credit risk models, including model development, validation, governance, and performance monitoring.
- NPE & Collateral Management: Weaknesses in the management of non-performing and forborne exposures, recovery and restructuring processes, as well as collateral valuation, monitoring, and legal enforceability.
- Credit Lifecycle Management: Deficiencies in credit granting, underwriting, approval, monitoring, and early warning processes throughout the lifecycle of a credit exposure.

While the finding type variable reflects the main topic of the findings, the number of topics reflects all those mentioned in the finding description.

The dataset is further enriched with bank-level variables drawn from Supervisory data. These include:

- The materiality of credit risk, calculated as the ratio between credit risk capital requirements and the total Capital Requirements
- A profitability measure of asset turnover, calculated as revenues over total assets.
- the SREP credit governance score (between 1 and 4, with 4 being the worst), reflecting supervisory assessments of risk management and governance frameworks.

All bank-level variables are measured one quarter prior to the issuance of the measure, unless otherwise specified.

To provide an overview of the sample, Annex 1 presents summary statistics and a description for the main variables used in the analysis (Tables A1-A3). Additional information on the distribution of supervisory measures across effectiveness categories, measure characteristics and finding types is reported in Figures A1–A3 in Annex 1.

5. Results

5.1 Drivers of effectiveness

Table 1 presents the results of the baseline binomial logistic regression, which examines the characteristics associated with the full remediation of supervisory measures. The dependent variable distinguishes between measures classified as fully effective and all other outcomes. Throughout the analysis, the estimated relationships are interpreted as associations rather than causal effects. Coefficients are reported in log-odds units.

Finding characteristics: Measures associated with findings covering a larger number of topics are significantly less likely to achieve full remediation, suggesting that more complex deficiencies are inherently harder to address. Such findings often require coordinated action across multiple functions, involve a broader set of stakeholders, and necessitate more extensive implementation efforts.

The type of deficiency also matters. Relative to Data, Reporting & IT findings, measures addressing Governance & Controls, Classification & Staging, Provisioning & Model Risk, and Credit Lifecycle Management exhibit a higher likelihood of full remediation. No statistically

significant difference is observed for NPE & Collateral Management findings. One possible explanation is that data and IT deficiencies typically require substantial system enhancements, data-quality improvements, and technology investments, whereas weaknesses in governance, credit processes, or provisioning practices can often be addressed through more targeted corrective actions.

The supervisory origin of a finding also matters. Compared with measures stemming from on-site inspections, those originating from the SREP or other JST activities exhibit lower remediation rates. A possible explanation is that on-site inspections generate more targeted and granular findings, facilitating the design and implementation of corrective actions.

Measure characteristics: Measure characteristics are also associated with remediation outcomes. Recommendations are more likely to achieve full remediation than obligations. This may reflect the fact that recommendations are often less complex to implement, making corrective action easier for supervised institutions.

Measure severity is negatively associated with remediation success. While low-severity measures do not differ significantly from those of moderate severity, high-severity (M3) and very high-severity (M4) measures are substantially less likely to achieve full remediation. This result is consistent with the notion that more severe findings tend to reflect deeper-rooted deficiencies that require more extensive and resource-intensive remediation efforts.

Bank characteristics: Bank characteristics are measured prior to the issuance of the supervisory measure.

Weaker governance, as proxied by a higher SREP Governance score, is associated with a lower likelihood of full remediation. This finding suggests that governance shortcomings may impair a bank's ability to implement supervisory measures effectively and sustain remediation efforts over time.

Further, in the baseline specification, banks with higher credit risk materiality and higher profitability are more likely to achieve full remediation. These results should, however, be interpreted with caution. While the governance effect remains significant when country fixed effects are included, the estimated effects of credit risk materiality and profitability weaken substantially. Once JST fixed effects are introduced, bank characteristics lose their significance, which is unsurprising as the JST fixed effect equals bank fixed effects. This suggests that part of the variation captured by the baseline estimates may reflect persistent

differences across supervised institutions, supervisory teams, or supervisory approaches, rather than independent effects of these bank-level characteristics.

Table 1: Logit Regression Results

Regressors		Coefficient	Std. Error	p-value
Intercept		-0.400*	0.204	0.050
Finding characteristics				
Number of topics		-0.142***	0.034	<0.001
Finding type (reference category: Data, Reporting & IT)	Governance & Controls	0.248**	0.093	0.007
	Classification & Staging	0.361***	0.096	<0.001
	Provisioning & Model Risk	0.264**	0.097	0.006
	NPE & Collateral Management	0.150	0.105	0.152
	Credit Lifecycle Management	0.224*	0.101	0.026
Process originator (reference category: OSI)	Horizontal	-0.495***	0.084	<0.001
	JST	-0.447***	0.070	<0.001
	SREP	-0.470***	0.104	<0.001
Measure characteristics				
Severity (reference category: moderate (M2))	Low severity (M1)	0.146	0.092	0.111
	High severity (M3)	-0.187***	0.055	<0.001
	Very high severity (M4)	-0.449***	0.128	<0.001
Legal enforceability (reference category: obligation)	Recommendation	0.402**	0.134	0.002
Macroeconomic and market controls				
GDP Growth		0.003	0.008	0.725
Unemployment rate		0.011	0.008	0.154
Herfindahl Index		-0.258***	0.028	<0.001
Bank characteristics				
Governance score		-0.105***	0.025	<0.001
Credit risk materiality		0.097***	0.028	<0.001
Asset turnover		0.068*	0.026	0.009
Year fixed effects (ref. year omitted)				
2017		0.282*	0.132	0.032
2018		0.257†	0.138	0.062
2019		0.431**	0.139	0.002
2020		0.331*	0.153	0.031
2021		0.032	0.140	0.082

Regressors	Coefficient	Std. Error	p-value
2022	0.365**	0.142	0.010
2023	0.628***	0.132	<0.001
2024	0.646***	0.139	<0.001
2025	0.164	0.286	0.567
Number of observations	6,969		
Null Deviance	9,526.6		
Residual Deviance	9,106.8		
AIC	9,164.8		
Fisher Scoring Iterations	4		
Significance	† p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001		

Time Trends: Finally, the year of measure creation is significantly associated with remediation outcomes. Relative to measures created in 2016, measures initiated in most subsequent years exhibit a higher probability of full remediation, with particularly pronounced effects for measures opened in 2019 and during 2023-2024. This pattern may reflect changes in supervisory practices, remediation processes, and follow-up intensity over time. It is also consistent with a stronger supervisory focus on the timely remediation of findings in the post-pandemic period. Since 2022, ECB Banking Supervision has strengthened its approach to findings and measures through the introduction of escalation frameworks, additional guidance for Joint Supervisory Teams (JSTs), and enhanced support for the management of complex remediation cases. In parallel, JSTs were encouraged to place greater emphasis on the follow-up of supervisory findings within their supervisory priorities and planning processes. However, these results should be interpreted with caution, as the sample includes only closed measures. More recent years may therefore be disproportionately represented by measures that were less complex and could be remediated within a shorter period.

Robustness analyses: Several robustness checks were conducted to assess the stability of the baseline results. First, the model was re-estimated alternatively including JST and country fixed effects. Second, an alternative binary logit model excluding ineffective outcomes was estimated. Third, a multinomial logit specification was estimated to distinguish explicitly between full remediation, partial remediation, and ineffective outcomes. Across all specifications, the results remain qualitatively unchanged and continue to provide strong support for the main findings regarding finding complexity, topic area, measure severity, and process originator. Detailed results are reported in Annex 2.

5.2 Drivers of measure duration

Table 2 reports the results of a Cox proportional hazards model explaining the time required to close fully effectively remediated supervisory measures i.e. a subset of all closed measures.

Finding characteristics: The number of topics addressed within a measure is negatively associated with the speed of closure. Measures covering multiple topics require significantly longer remediation periods, suggesting that more complex findings are inherently more difficult to resolve. This result is consistent with the notion that addressing several interconnected deficiencies often requires broader organisational changes, greater coordination across business areas, and more extensive supervisory follow-up.

The nature of the underlying deficiency also plays an important role. Measures related to Data, Reporting & IT exhibit the longest remediation periods, while all other finding categories are associated with significantly faster closure. In particular, Governance & Controls, NPE & Collateral Management, and Credit Lifecycle Management findings tend to be resolved more quickly. Remediation actions involving data infrastructures, reporting frameworks, and IT systems frequently require substantial investments, system development, and implementation efforts across multiple functions. By contrast, corrective actions targeting governance arrangements, credit processes, provisioning practices, or collateral management can often be implemented within shorter timeframes, resulting in faster remediation.

The results indicate that the originator of the supervisory process is an important determinant of remediation speed. Relative to measures stemming from on-site inspections (OSI), measures originating from the SREP, Horizontal analyses, and JST activities exhibit significantly lower closure hazards, implying longer remediation periods. One possible explanation is that findings arising from broader supervisory assessments often relate to more structural or cross-cutting deficiencies, which may require more extensive remediation efforts before closure can be granted.

Measure characteristics:

Measure severity is also strongly associated with remediation duration. Compared with moderately severe measures (M2), low-severity measures (M1) are remediated more quickly, while high-severity (M3) and very high-severity (M4) measures remain open for longer. This pattern is consistent with the view that more severe findings typically reflect more significant

underlying weaknesses and therefore require more complex corrective actions and supervisory verification.

By contrast, legal enforceability does not appear to affect remediation timelines. Recommendations do not differ significantly from legally binding obligations in terms of closure speed, suggesting that, conditional on other measure characteristics, the formal legal status of a measure is not a key determinant of how quickly remediation is achieved.

Bank characteristics: Among bank-level factors, higher credit risk materiality and greater profitability are associated with faster remediation. This may indicate that institutions for which credit risk is a more central area of activity, and those with stronger financial performance, are better positioned to allocate resources to the implementation of corrective actions. However, these relationships weaken substantially once country and JST fixed effects are introduced, suggesting that part of the observed association reflects persistent differences across supervised institutions or supervisory teams rather than independent effects of these bank characteristics.

By contrast, weaker governance, as measured by a higher SREP Governance score, is associated with faster closure among fully effective measures. This result should be interpreted cautiously. As shown in Section 5.1, weaker-governance banks are less likely to achieve full remediation in the first place. Since the duration analysis is conditional on successful remediation, the result may partly reflect selection effects, whereby only a subset of measures with a relatively high likelihood of remediation reaches closure in weaker-governance institutions. It may also reflect supervisory prioritisation, with both banks and supervisors focusing remediation efforts on a smaller number of measures where successful completion is considered most achievable.

Table 2: Cox proportional hazards model

Regressors		Hazard Ratio	Change in Hazard Rate	Change in Time to Closure	p-value
Finding characteristics					
Number of topics		0.84***	16% lower	19% longer	<0.001
Finding type (reference category: Data, Reporting & IT)	Governance & Controls	1.39***	39% higher	28% shorter	<0.001
	Classification & Staging	1.19**	19% higher	16% shorter	0.009

Regressors		Hazard Ratio	Change in Hazard Rate	Change in Time to Closure	p-value
	Provisioning & Model Risk	1.22**	22% higher	18% shorter	0.003
	NPE & Collateral Management	1.35***	35% higher	26% shorter	<0.001
	Credit Lifecycle Management	1.29***	29% higher	23% shorter	<0.001
Process originator (reference category: OSI)	Horizontal	0.77***	23% lower	29% longer	<0.001
	JST	0.90*	10% lower	11% longer	0.032
	SREP	0.66***	34% lower	52% longer	<0.001
Measure characteristics					
Severity (reference category: moderate (M2))	Low severity (M1)	1.28***	28% higher	22% shorter	<0.001
	High severity (M3)	0.91**	9% lower	10% longer	0.008
	Very high severity (M4)	0.82*	18% lower	21% longer	0.042
Legal enforceability (reference category: obligation)	Recommendation	1.05	5% higher	5% shorter	0.623
Macroeconomic and market controls					
GDP growth		1.01*	1% higher	1% shorter	0.024
Unemployment rate		0.99*	1% lower	1% longer	0.044
Herfindahl index		0.96*	4% lower	4% longer	0.036
Bank characteristics					
Governance score		1.13***	13% higher	11% shorter	<0.001
Credit risk materiality		1.06**	6% higher	6% shorter	0.006
Asset turnover		1.04*	4% higher	4% shorter	0.026
Year fixed effects (ref. year 2016)					
2017		0.65***	35% lower	54% longer	<0.001
2018		0.46***	54% lower	115% longer	<0.001
2019		0.43***	57% lower	134% longer	<0.001
2020		0.53***	47% lower	89% longer	<0.001
2021		0.58***	42% lower	72% longer	<0.001
2022		0.62***	38% lower	61% longer	<0.001
2023		0.66***	34% lower	50% longer	<0.001
2024		0.97	3% lower	3% longer	0.742

Regressors	Hazard Ratio	Change in Hazard Rate	Change in Time to Closure	p-value
2025	1.80**	80% higher	44% shorter	0.003
Observations	3,641			
Events	3,641			
Concordance (C-index)	0.617			
Significance	† p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001			

Robustness analyses: One potential concern is that the estimated relationships may be influenced by unobserved differences across supervised institutions, supervisory teams or national supervisory environments. To address this, the Cox model was re-estimated including alternatively country and JST fixed effects. In particular, the estimated effects of finding complexity, finding type, process originator and measure severity remain stable across specifications. By contrast, the estimated effects of credit risk materiality, profitability and governance weaken substantially once JST fixed effects are included. This suggests that part of the variation attributed to bank characteristics in the baseline specification reflects persistent differences across supervised institutions or supervisory teams.

In addition, the proportional hazards diagnostics reported in Annex 3 indicate violations of the proportional hazards assumption for several covariates and for the model globally. Hence, as an additional robustness check, Annex 4 reports estimates from a Weibull Accelerated Failure Time (AFT) model, which does not rely on this assumption.

The AFT specification confirms the main findings of the Cox model. Measures associated with a larger number of findings exhibit significantly longer remediation durations, while findings related to Governance & Controls, Classification & Staging, Provisioning & Model Risk, NPE & Collateral Management, and Credit Lifecycle Management are closed more rapidly than findings in the reference category. SREP-originated measures are estimated to remain open approximately 4% longer than otherwise comparable measures, whereas measures originating from Horizontal functions and JSTs also exhibit modestly longer remediation durations. Weaker governance scores and higher credit risk materiality are associated with significantly shorter closure times, while legal enforceability continues to have no statistically significant effect. Although some individual coefficients differ in magnitude from those obtained in the Cox model, the overall pattern of results remains unchanged.

Taken together, the consistency of the results across alternative specifications, including the AFT model and the additional bank-level controls, provides further evidence that the main conclusions are robust.

6. Conclusion

This paper provides a granular, outcome-oriented assessment of supervisory effectiveness by focusing on the remediation of supervisory measures in credit risk. Using a novel application of LLM to confidential supervisory data, it constructs a measure-level indicator of effectiveness. Further, using the subset of fully effectively closed measures, a duration analysis is conducted. Both analyses assess drivers using measure characteristics and bank-specific factors providing a comprehensive narrative about their influence on remediation outcomes.

Finding issue: The results show that finding complexity is an important determinant of supervisory outcomes. Measures addressing a larger number of topics are less likely to achieve full remediation and take substantially longer to close. In addition, Data, Reporting & IT findings emerge as particularly challenging to remediate. Relative to other finding types, such measures display both lower remediation rates and longer remediation horizons, highlighting the implementation challenges associated with data quality, reporting frameworks and IT infrastructures. Findings originating from on-site inspections outperform those stemming from broader supervisory activities, such as the SREP and ongoing JST assessments, both in terms of remediation rates and closure times.

Measure Characteristics: Supervisory measure characteristics are strongly associated with both effectiveness and remediation speed. More severe measures are less likely to achieve full remediation and require longer implementation periods. Recommendations are associated with a higher likelihood of full remediation, although legal enforceability does not appear to influence remediation speed. Taken together, these findings suggest that the nature of the underlying deficiency, the supervisory tool employed, and the complexity of implementation are important correlates of supervisory outcomes.

Bank characteristics: Bank characteristics play a more nuanced role. More profitable institutions and banks with greater credit-risk materiality display stronger remediation outcomes in the baseline specifications, although these relationships weaken once JST-specific heterogeneity is considered. By contrast, weaker governance is consistently associated with a lower likelihood of full remediation. At the same time, among measures that are fully remediated, weaker-governance institutions tend to achieve closure more quickly, a

result that may reflect selection effects or supervisory prioritisation rather than a direct governance effect on remediation speed.

Time and Temporal Trends: The results indicate that measures initiated in most years after 2016 are more likely to achieve full remediation, with particularly strong improvements observed in recent years. This pattern is consistent with an increased supervisory focus on remediation and follow-up, including the introduction of dedicated escalation frameworks and supervisory guidance supporting the closure of findings.

Policy Implications: The findings offer several actionable insights for supervisors and policymakers:

Simplifying Complex Measures: The greater effectiveness of recommendations and on-site measures may suggest that simplifying the implementation of measures could improve compliance. Supervisors should focus on feasibility, particularly for less well-resourced or poorly governed institutions.

Tailoring Supervisory Strategies: Supervisors should account for bank-specific characteristics, such as governance, when designing and enforcing measures. The findings suggest that remediation strategies may need to be tailored to the characteristics and implementation capacity of supervised institutions.

More broadly, the paper demonstrates how artificial intelligence can support the systematic evaluation of supervisory outcomes by transforming large volumes of unstructured supervisory information into empirically usable data. While the analysis identifies associations rather than causal effects and relies on documented remediation captured in supervisory narratives, it provides a practical framework for monitoring supervisory effectiveness at scale and contributes new evidence to the growing policy discussion on how supervisory success should be measured.

References

Adrian, T., Moretti, M., Carvalho, A., Kyong Chon, H., Seal, K. Melo, F., and Surti, J. (2023) “Good Supervision: Lessons from the Field”, IMF Working paper WP/23/181

Agarwal, S., Lucca, D., Seru, A. and Trebbi, F. (2014), "Inconsistent regulators: evidence from banking", Quarterly Journal of Economics, Vol. 129, No 2, pp. 889-938.

Bank for International Settlements (2024), "Large language models: a primer for economists", BIS Quarterly Review, December.

Basel Committee on Banking Supervision (2025), "Lessons on supervisory effectiveness – a literature review", BCBS Working Papers, No 45, Bank for International Settlements, July.

Berger, A. and Davies, S. (1998), "The information content of bank examinations", Journal of Financial Services Research, Vol. 14, No 2, pp. 117-144.

Berger, A. N., Davies, S. M. and Flannery, M. J. (2000). ‘Comparing market and supervisory assessments of bank performance: Who knows what when?’, Journal of Money, Credit and Banking, 32(3), pp. 641–667. doi: 10.2307/2601200.

Berger, A. N., Cai, J., Roman, R. and Sedunov, J. (2022). ‘Supervisory enforcement actions against banks and systemic risk’, Journal of Banking and Finance, 140, 106222. doi: 10.1016/j.jbankfin.2021.106222.

Bonfim, D., Cerqueiro, G. Degryse, H. and Ongena, S. (2023). ‘On-site inspecting zombie lending’, Management Science, 69(5), pp. 2547–3155. doi: 10.1287/mnsc.2022.4452.

Caiazza, S., Cotugno, M., Fiordelisi, F. and Stefanelli, V. (2018). ‘The spillover effect of enforcement actions on bank risk-taking’, Journal of Banking and Finance, 91, pp. 146–159. doi: 10.1016/j.jbankfin.2018.04.008.

Correa, R., Garud, K., Londono, J.M. and Mislang, N. (2021), "Sentiment in central banks' financial stability reports", Review of Finance, Vol. 25, No 1, pp. 85-120.

Delis, M. D., Staikouras, P. K. and Tsoumas, C. (2016). ‘Formal enforcement actions and bank behaviour’, Management Science 63(4), pp. 959–987. doi: 10.1287/mnsc.2015.2343.

Eisenbach, T.M., Lucca, D.O. and Townsend, R.M. (2016), "The economics of bank supervision", NBER Working Paper, No 22201, National Bureau of Economic Research.

Eisenbach, T.M., Lucca, D.O. and Townsend, R.M. (2022), "Resource allocation in bank supervision: trade-offs and outcomes", Journal of Finance, Vol. 77, No 3, pp. 1685-1736.

Fell, J., Gardó, S., Kellner, D., Klaus, B., Lang, J.H., Nagy, L., Vendrell, P., Rusnák, M., Wendelborn, J. and Wredenborg, S. (2026), "From dictionaries to AI: a new era in sentiment analysis for financial stability", Financial Stability Review, Special Feature, European Central Bank, May.

Gentzkow, M., Kelly, B. and Taddy, M. (2019), "Text as data", Journal of Economic Literature, Vol. 57, No 3, pp. 535-574.

Goldsmith-Pinkham, P., Hirtle, B. and Lucca, D. (2016), "Parsing the content of bank supervision", Federal Reserve Bank of New York Staff Reports, no. 770.

Hansen, S., McMahon, M. and Prat, A. (2018), "Transparency and deliberation within the FOMC: a computational linguistics approach", Quarterly Journal of Economics, Vol. 133, No 2, pp. 801-870.

Hirtle, B., Kovner, A. and Plosser, M. (2020), "The impact of supervision on bank performance", Journal of Finance, Vol. 75, No 5, pp. 2765-2808.

Huang, A.H., Wang, H. and Yang, Y. (2023), "FinBERT: a large language model for extracting information from financial text", Contemporary Accounting Research, Vol. 40, No 2, pp. 806-841.

Korinek, A. (2023), "Generative AI for economic research: use cases and implications for economists", Journal of Economic Literature, Vol. 61, No 4, pp. 1281-1317.

Loughran, T. and McDonald, B. (2011), "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks", Journal of Finance, Vol. 66, No 1, pp. 35-65.

Passalacqua, A., Angelini, P., Lotti, F. and Soggia, G. (2021). The real effects of bank supervision: evidence from on-site bank inspections. Bank of Italy, Temi di Discussione (working paper), No. 1349. Available at: papers.ssrn.com/sol3/papers.cfm?abstract_id=3705558.

Annex

Technical Annex A - Effectiveness prompt

This annex documents the large language model (LLM) classification protocol used to construct the supervisory effectiveness variable.

Model specification

- Model: GPT-4.1
- Provider: OpenAI
- Temperature: 0.15
- Number of independent classifications per observation: 10
- Aggregation rule: The final effectiveness label was determined by majority voting across the ten independent model runs. Cases without a clear majority were manually reviewed by the research team.

Classification task

We instructed the LLM to evaluate whether the closure narrative provided evidence that the supervisory measure had successfully remediated the underlying deficiency identified by supervisors. Effectiveness was assessed according to the following definition:

"Supervisory effectiveness means that the supervisory toolkit and powers are used to ensure that banks fully remediate identified deficiencies."

We instructed the model to assign one of three categories:

- **Ineffective**: The closure narrative provides little or no evidence that the identified deficiency was remediated.
- **Somewhat effective**: The closure narrative indicates partial remediation or progress toward remediation, but material deficiencies remain or evidence is incomplete.
- **Effective**: The closure narrative provides evidence that the identified deficiency was fully remediated and the supervisory objective was achieved.

Input data

For each observation, we provided the model with:

- Measure title/header
- Measure description

- Closure rationale

Quality assurance and validation

Several validation procedures were implemented:

1. Expert review of selected subsamples.
2. Iterative prompt refinement based on supervisory expert feedback.
3. Ten independent classification runs per observation, aggregated using majority voting.
4. Robustness analysis using an alternative model (o4-mini), yielding an agreement rate of approximately 90%.
5. Manual inspection of cases displaying disagreement across iterations.

Limitations

The effectiveness variable should be interpreted as an LLM-assisted assessment of remediation outcomes based on information contained in supervisory closure narratives. While the classification protocol was designed to maximise consistency through detailed instructions, repeated model runs, majority voting, and expert validation, any assessment of effectiveness inevitably involves an element of interpretation. The resulting measure therefore reflects a structured expert-informed evaluation of the evidence contained in closure documentation rather than a mechanically observable outcome.

For reproducibility, the complete system prompts are available from the corresponding author upon request.

Technical annex B – Finding's topic classification

This annex documents the large language model (LLM) workflow used to classify supervisory findings and remediation measures into substantive credit-risk topics.

The process consisted of two steps:

- Taxonomy Development Stage: inductive identification of supervisory remediation themes from the text of findings and measures.
- Taxonomy Classification Stage: systematic assignment of individual measures to taxonomy categories using step 1 classification framework.

B.1 - Taxonomy development stage

The first step was designed to generate a parsimonious and scalable taxonomy of supervisory findings and remediation measures based on the deficiencies identified by supervisors and substantive remediation objective underlying each measure.

Input

For each supervisory case, the model received:

- Finding description
- Measure description

Model Configuration

- Model: GPT-4.1
- Provider: OpenAI
- Temperature: 0.00

Quality assurance and validation

To assess robustness, the taxonomy-generation exercise was repeated under multiple prompt formulations and applied to both the full sample and temporal subsamples.

The final prompt specification was executed five times. The resulting taxonomies exhibited a high degree of consistency regarding the principal thematic categories and their hierarchical structure.

Following generation, supervisory experts reviewed the proposed taxonomy and refined category definitions, resolved overlaps, and consolidated conceptually similar categories. The final taxonomy therefore reflects an iterative human-in-the-loop process combining LLM-assisted thematic discovery with expert validation.

B2 - Taxonomy Classification Stage

In the second step, the refined taxonomy is embedded in a second prompt where we ask the LLM to assign each measure to the most appropriate taxonomy category.

Measures were classified according to the primary credit-risk topic being remediated.

Input

For each observation, the model received:

- Measure title/header

- Finding description
- Measure description

Model Configuration

- Model: GPT-4.1
- Provider: OpenAI
- Temperature: 0.15

Classification Principle

The model was instructed to classify measures based on the primary credit-risk topic being remediated, rather than the remediation instrument used.

Accordingly, measures referring to policies, procedures, methodologies or frameworks were classified according to the substantive risk area addressed, rather than automatically being allocated to the policy category.

Multiple tags were permitted only when a measure contained distinct remediation actions spanning multiple macro-topic areas.

Quality assurance and validation

Several measures were implemented to enhance consistency and reliability:

- Finding and measure descriptions were evaluated jointly;
- Predefined inclusion and exclusion criteria were embedded in the prompt;
- The model was instructed to identify the primary remediation objective;
- Category assignments based solely on policy or documentation language were discouraged;
- Short evidence-based rationales were required for each classification.

The classification prompt was developed iteratively. Multiple prompt formulations were tested on randomly selected subsamples, and outputs were reviewed by supervisory experts. Revisions continued until classifications were stable and conceptually coherent across testing samples.

Following full-sample classification, experts reviewed category distributions and representative observations to identify remaining overlaps, redundancies, or conceptual ambiguities.

This review indicated that several categories captured closely related remediation objectives and could be consolidated into broader supervisory themes without a material loss of information.

As a result, the original ten-category taxonomy was aggregated into six higher-level analytical categories used in the empirical analysis:

Original Category	Final Category
A. Credit Risk Governance, Oversight & Internal Controls	Governance & Controls
B. Policy Framework & Documentation	Governance & Controls
J. Risk Appetite, Portfolio Strategy & Group Risk Management	Governance & Controls
D. Credit Risk Classification, Identification & Staging	Classification & Staging
C. Data, Reporting & IT Infrastructure	Data, Reporting & IT
G. Provisioning, Impairment & Model Risk	Provisioning & Model Risk
E. NPL, NPE & Forbearance Strategy and Management	NPE & Collateral Management
F. Collateral Management & Valuation	NPE & Collateral Management
H. Credit Monitoring & Early Warning	Credit Lifecycle Management
I. Credit Granting, Underwriting & Approval Process	Credit Lifecycle Management

Consequently, the final classification framework should be understood as the outcome of a structured human-machine collaboration in which LLM-assisted classification was complemented by iterative prompt refinement and expert validation.

Limitations

The resulting classifications depend on the information contained in supervisory findings and measure descriptions. As with other LLM-based annotation approaches, outputs may be sensitive to prompt design and model choice. To mitigate these risks, the taxonomy and classification rules were developed through repeated testing, expert review, and iterative refinement.

For reproducibility, the complete system prompts are available from the corresponding author upon request.

Annex 1 – Descriptive statistics

Table A1: Distribution of Measure-Level Categorical Variables

Categorical Variable	Category	Percentage
Effectiveness	Effective	55.25%
	Ineffective	1.51%
	Somewhat effective	43.24%

Measure severity	M1: low	8.80%
	M2:moderate	47.29%
	M3: high	39.12%
	M4: very high	4.80%
Process originator	Horizontal	11.35%
	JST	18.24%
	OSI	60.65%
	SREP	9.76%
Legal enforceability	Obligation	6.02%
	Recommendation	93.98%
Finding topic	Governance & Controls	21.39%
	Classification & Staging	20.56%
	Provisioning & Model Risk	18.92%
	NPE & Collateral Management	13.06%
	Credit Lifecycle Management	15.72%
	Data, Reporting & IT findings	10.34%

Table A2: Summary Statistics of Bank-Level Variables

Variable	Min	Q1	Median	Mean	Q3	Max	Coverage
Number of topics	1.000	1.000	1.000	1.535	2.000	10.000	100%
Credit risk materiality	0.310	0.814	0.8525	0.842	0.885	0.972	92%
Asset turnover	-0.017	0.008	0.013	0.014	0.019	0.051	92%
SREP governance score	2.000	2.600	3.000	2.909	3.000	4.000	89%

Table A3: Distribution of Measures by year of Creation

Year of Measure Creation	Number of Measures
2016	396
2017	1051
2018	860
2019	678
2020	681
2021	915
2022	685
2023	1407
2024	998
2025	435

The distribution of measures over time is relatively balanced across the sample period, with observations ranging from 396 to 1407 per year between 2016 and 2025. In addition, the restriction to closed measures may introduce a selection effect. For more recent years, only measures with relatively short implementation horizons are observed, which are likely to be less complex and easier to remediate.

The distribution of the dependent variable shows a relatively balanced split between fully effective measures (55.25%) and somewhat effective measures (43.24%), while only a small share (1.51%) are classified as ineffective. This indicates that outright non-remediation is rare in the sample, with most measures achieving at least partial implementation.

Regarding supervisory measure characteristics, most observations correspond to moderate-severity (M2) and high-severity (M3), accounting for around 47.29% and 39.12% of the sample respectively, while low-severity (M1) and very high-severity (M4) measures represent approximately 8.80% and 4.80%. In terms of supervisory origin, measures stemming from on-site inspections (OSI) dominate the sample (60.65%), followed by JST activities (18.24%) and horizontal exercises (11.35%), while SREP-related measures account for a smaller share (9.76%). With respect to legal enforceability, most measures take the form of non-binding recommendations (93.98%), whereas legally binding obligations account for only 6.02% of observations.

Turning to bank-level characteristics, the summary statistics indicate substantial heterogeneity across institutions. Credit risk materiality exhibits relatively high levels across the sample, with a mean of 84.2% and values ranging from 31.0% to 97.2%, suggesting significant variation in the importance of credit risk across banks' risk profiles. In terms of profitability, asset turnover, capturing the efficiency with which banks generate income from their asset base, has a mean of 1.4% and ranges from -1.7% to 5.1%, indicating considerable differences in business models and operational performance. Finally, the SREP governance score is a supervisory assessment of the banks' internal governance and risk management framework. It may take values from 1 to 4 with 1 being the best. In the sample it ranges from 2 to 4, with a median of 3, suggesting that most banks in the sample are assessed as having moderate to satisfactory governance frameworks.

The number of topics reflected within findings also varies across institutions, with an average of 1.5 topics and a range from 1 to 10, pointing to differences in supervisory focus and different levels of complexity.

Figure A1: Distribution of effectiveness outcomes by legal enforceability

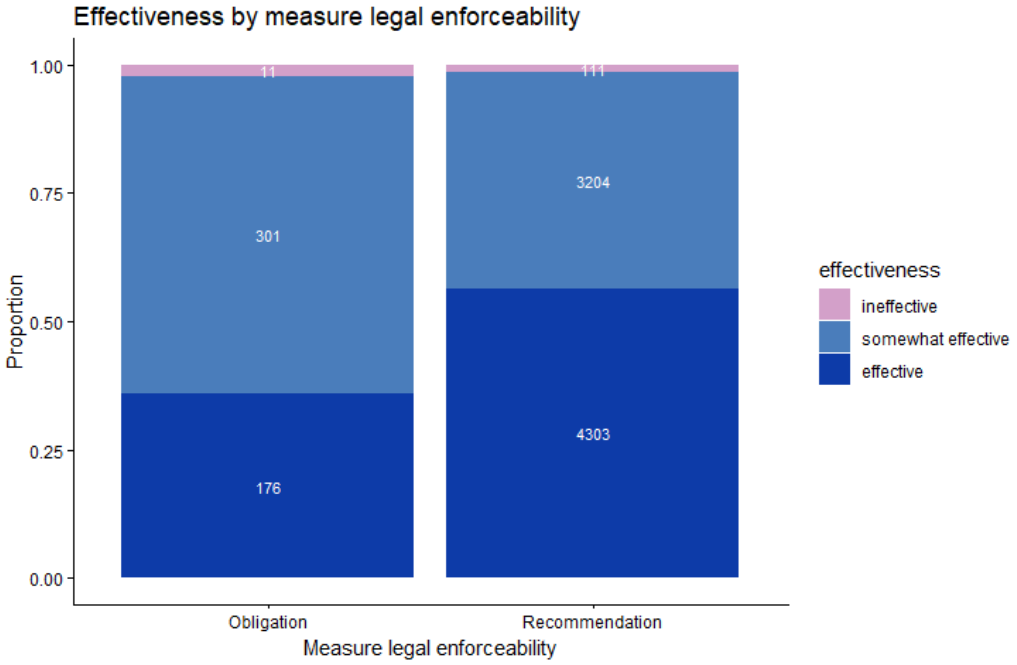


Figure A2: Distribution of effectiveness outcomes by source of the finding

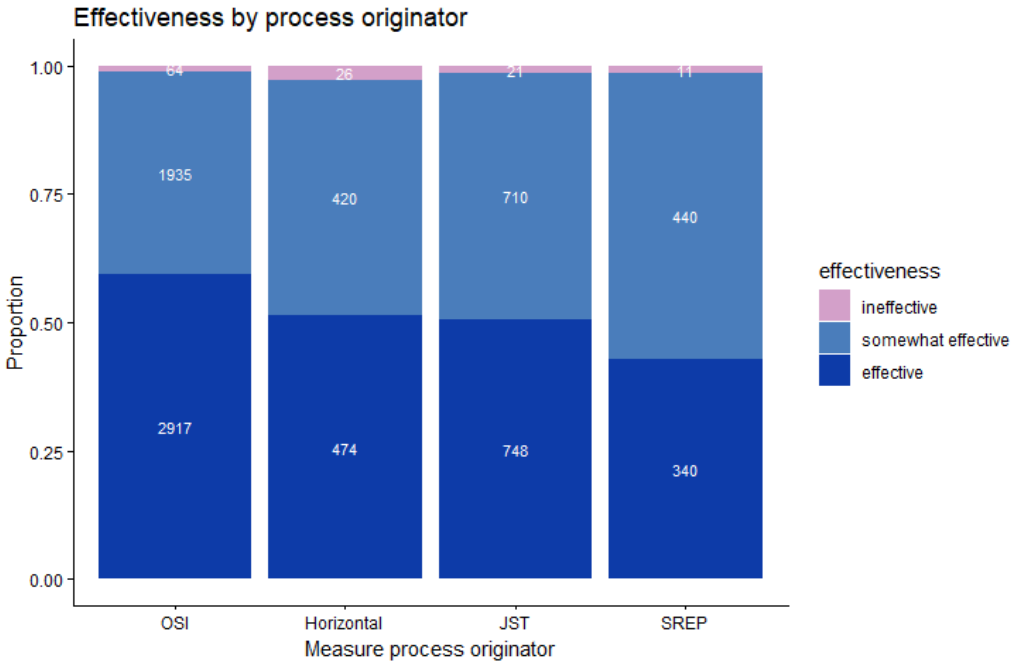
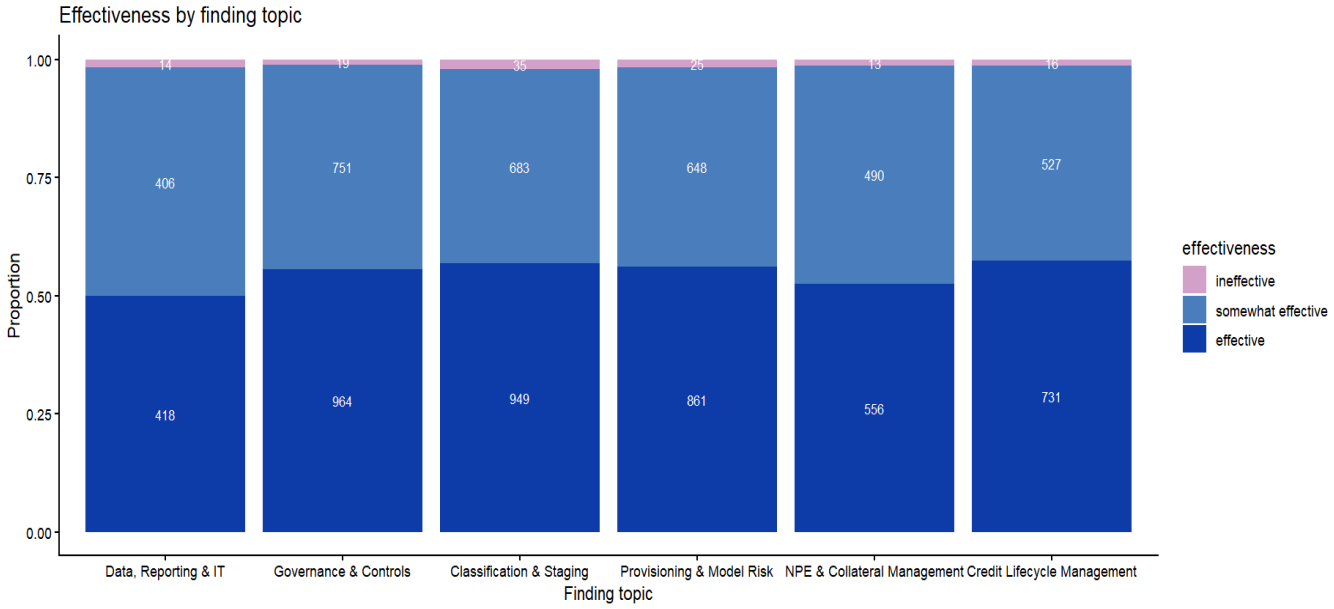


Figure A3: Distribution of effectiveness outcomes by type of finding issue



Annex 2 – Drivers of effectiveness: robustness checks

Table A4: Binomial logit augmented with JST and country fixed effects

Regressors		Country FE	JST FE
Finding characteristics			
Number of topics		*-0.122****	*-0.107***
Finding type (ref. Data, Reporting & IT)	*Governance & Controls*	*0.256***	*0.275***
	Classification & Staging	*0.340****	*0.372****
	Provisioning & Model Risk	*0.254***	*0.269***
	NPE & Collateral Management	*0.137*	*0.092*
	Credit Lifecycle Management	*0.199*	*0.247**
Process originator (ref. OSI)	*Horizontal*	*-0.521****	*-0.537****
	JST	*-0.449****	*-0.374****
	SREP	*-0.499****	*-0.481****
Measure characteristics			
Severity (ref. moderate, M2)	*Low severity (M1)*	*0.122*	*0.084*
	High severity (M3)	*-0.187****	*-0.168***
	Very high severity (M4)	*-0.484****	*-0.460****
Legal enforceability (ref. obligation)	*Recommendation*	*0.415***	*0.335**

Regressors	*Country FE*	*JST FE*
Macroeconomic and market controls		
GDP growth	*0.005*	*-0.007*
Unemployment rate	*0.027*	*0.012*
Herfindahl index	*0.205*	*0.025*
Bank characteristics		
Governance score	*-0.107***	*-0.015*
Credit risk materiality	*0.079***	*-0.029*
Asset turnover	*0.029*	*0.058*
***Year fixed effect** (ref. 2016)*		
2017	*0.293**	*0.249.*
2018	*0.213*	*0.210*
2019	*0.407***	*0.402**
2020	*0.320.*	*0.148*
2021	*-0.006*	*-0.017*
2022	*0.296.*	*0.249*
2023	*0.651****	*0.602****
2024	*0.596****	*0.546****
2025	*0.086*	*0.190*
Observations	*6,918*	*6,918*
Fixed effects	*Country*	*JST code*
AIC	*9,110.9*	*8,955.1*
Significance	† p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001	

Table A5: Binomial regression excluding ineffective measures

Regressors		Coefficient	Std. Error	p-value
Intercept		-0.303	0.206	0.142
Finding characteristics				
Number of topics		-0.139***	0.034	<0.001
Finding type (reference category: Data, Reporting & IT)	Governance & Controls	0.242*	0.094	0.010
	Classification & Staging	0.381***	0.097	<0.001
	Provisioning & Model Risk	0.270**	0.098	0.006
	NPE & Collateral Management	0.147	0.105	0.163

Regressors		Coefficient	Std. Error	p-value
Process originator (reference category: OSI)	Credit Lifecycle Management	0.224*	0.101	0.027
	Horizontal	-0.463***	0.085	<0.001
	JST	-0.442***	0.070	<0.001
	SREP	-0.465***	0.104	<0.001
Measure characteristics				
Severity (reference category: moderate (M2))	Low severity (M1)	0.126	0.092	0.172
	High severity (M3)	-0.176**	0.056	0.002
	Very high severity (M4)	-0.475***	0.128	<0.001
Legal enforceability (reference category: obligation)	Recommendation	0.403**	0.135	0.003
Macroeconomic and market controls				
GDP Growth		-0.001	0.008	0.952
Unemployment rate		0.008	0.008	0.336
Herfindahl Index		-0.232***	0.029	<0.001
Bank characteristics				
Governance score		-0.106***	0.025	<0.001
Credit risk materiality		0.083**	0.028	0.003
Asset turnover		0.061*	0.026	0.019
Year fixed effects (ref. year omitted)				
2017		0.272*	0.134	0.043
2018		0.243†	0.140	0.083
2019		0.383**	0.141	0.007
2020		0.266†	0.155	0.087
2021		0.005	0.142	0.974
2022		0.306*	0.144	0.034
2023		0.556***	0.134	<0.001
2024		0.557***	0.141	<0.001
2025		0.062	0.286	0.829

Number of observations	6,803
Null Deviance	9,341.6
Residual Deviance	8,965.2
AIC	9,023.2
Fisher Scoring Iterations	4
Significance	† p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001

Table A6: Multinomial regression

Somewhat effective vs effective		**RRR**	**Magnitude (%)**	**p-value**
		Finding characteristics		
Number of topics		1.15***	+15.1%	<0.001
Finding type (reference category: Data, Reporting & IT)	Governance & Controls	0.79*	−21.3%	0.011
	Classification & Staging	0.68***	−31.6%	<0.001
	Provisioning & Model Risk	0.77**	−23.5%	0.006
	NPE & Collateral Management	0.86	−13.6%	0.165
	Credit Lifecycle Management	0.80*	−20.2%	0.026
Process originator (reference category: OSI)	Horizontal	1.59***	+59.1%	<0.001
	JST	1.56***	+55.8%	<0.001
	SREP	1.60***	+59.7%	<0.001
		Measure characteristics		
Severity (reference category: moderate (M2))	Low severity	0.88	−11.8%	0.174
	High severity	1.19**	+19.0%	0.002
	Very high severity	1.60***	+60.1%	<0.001
Legal enforceability (reference category: obligation)	Recommendation	0.67**	−32.9%	0.003
		Macroeconomic environment		
GDP growth		1.00	+0.1%	0.886
Unemployment rate		0.99	−0.7%	0.347
Herfindahl index		1.26***	+25.9%	<0.001
		Bank characteristics		
Governance score		1.11***	+10.9%	<0.001
Credit risk materiality		0.92**	−8.1%	0.003
Asset turnover		0.94*	−5.9%	0.020
		Year fixed effects (vs 2016)		
2017		0.77*	−23.4%	0.047
2018		0.79†	−20.6%	0.099
2019		0.69**	−31.5%	0.007
2020		0.78	−22.2%	0.106

Somewhat effective vs effective	RRR	Magnitude (%)	p-value
2021	0.99	−0.5%	0.969
2022	0.74*	−26.0%	0.037
2023	0.58***	−42.3%	<0.001
2024	0.58***	−42.3%	<0.001
2025	0.94	−5.7%	0.837
Significance	† p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001		

Annex 3 – Test for cox hazard proportionality assumptions

Table A7: Proportionality assumption test

Regressor	χ^2	df	p-value	PH Assumption
Number of topics	16.04	1	<0.001	Violated
Finding type	12.48	5	0.029	Violated
Measure severity	49.73	3	<0.001	Violated
Measure process originator	109.94	3	<0.001	Violated
Legal enforceability	25.60	1	<0.001	Violated
GDP growth	21.39	1	<0.001	Violated
Unemployment rate	0.28	1	0.598	Not rejected
Herfindahl index	7.12	1	0.008	Violated
Governance score	19.18	1	<0.001	Violated
Credit risk materiality	26.03	1	<0.001	Violated
Asset turnover	1.74	1	0.187	Not rejected
Year fixed effects	212.79	9	<0.001	Violated
Global test	**390.88**	**28**	**<0.001**	**Violated**

Schoenfeld residual tests indicate violations of the proportional hazards assumption for several covariates and for the model globally. Visual inspection of the residual plots suggests that most departures from proportionality are gradual rather than indicative of abrupt changes in effect direction. To assess robustness, a Weibull accelerated failure time model was estimated. The AFT results are qualitatively similar to those obtained from the Cox specification, indicating that the main findings are not driven by the proportional hazards assumption.

Annex 4 – Drivers of duration: robustness checks

4.1 JST and country fixed effects

Table A8: Cox augmented with JST and country fixed effects

Regressors		Country fixed effects (HR)	JST fixed effects (HR)
Finding characteristics			
Number of topics		0.837***	0.850***
Finding type (reference category: Data, Reporting & IT)	Governance & Controls	1.411***	1.417***
	Classification & Staging	1.124†	1.155*
	Provisioning & Model Risk	1.159*	1.208**
	NPE & Collateral Management	1.339***	1.319***
	Credit Lifecycle Management	1.251**	1.289***
Process originator (reference category: OSI)	Horizontal	0.707***	0.715***
	JST	0.857**	0.789***
	SREP	0.602***	0.521***
Measure characteristics			
Severity (reference category: moderate (M2))	Low severity (M1)	1.246***	1.226***
	High severity (M3)	0.907**	0.872***
	Very high severity (M4)	0.806*	0.871
Legal enforceability (reference category: obligation)	Recommendation	1.090	1.140
Macroeconomic and market controls			
GDP growth		1.012†	1.005
Unemployment rate		0.944**	0.981
Herfindahl index		0.877	1.068
Bank characteristics			
Governance score		1.107***	1.187***
Credit risk materiality		1.042†	1.011
Asset turnover		0.976	0.992
Year fixed effects (ref. year 2016)			
2017		0.752**	0.702***
2018		0.524***	0.461***
2019		0.447***	0.425***
2020		0.546***	0.476***
2021		0.649***	0.643***

Regressors	Country fixed effects (HR)	JST fixed effects (HR)
	Finding characteristics	
2022	0.637***	0.619***
2023	0.707***	0.649***
2024	0.969	0.849
2025	1.625*	2.269***
Observations	3,744	3,744
Events	3,744	3,744
Concordance	0.631	0.669
LR χ^2	611.1	1,078.0
Wald χ^2	617.8	1,093.0
Score χ^2	634.2	1,164.0
Significance	† p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001	

4.2 Accelerated Failure Time results

Table A9 reports estimates from a Weibull accelerated failure time model with JST-clustered standard errors and country fixed effects. Notably, unlike the Cox model, in AFT a positive (negative) coefficient implies longer (shorter) duration.

The results broadly confirm the findings from the Cox model. Measures associated with a larger number of findings take longer to close, while findings related to Governance & Controls, Classification & Staging, Provisioning & Model Risk, NPE & Collateral Management, and Credit Lifecycle Management are resolved significantly faster than the reference category (Data, Reporting & IT).

Regarding measure characteristics, SREP-originated measures remain open significantly longer than measures originating from OSIs, with expected closure times approximately 4% longer. Measures classified as low severity (M1) are resolved more quickly, whereas high severity measures (M3) exhibit longer remediation durations. Legal enforceability continues to have no statistically significant effect on closure times.

Bank characteristics also exert a meaningful influence on remediation duration. Worse governance scores are associated with significantly shorter closure times, while greater credit risk materiality is likewise linked to faster remediation. In short, this specification supports the results from the Cox model reported in the result section. Moreover, by using Weibull accelerated failure time model, which does not rely on proportional hazard assumptions, and supports the results of the Cox model, we resolve some concerns on the validity of previous assumptions.

Table A9: Robustness check – AFT

Regressors		Time Ratio	Change in Hazard Rate	Change in Time to Closure	p-value
Finding characteristics					
Number of topics		1.02***	2% lower	2% longer	<0.001
Finding type (reference category: Data, Reporting & IT)	Governance & Controls	0.97***	3% higher	3% shorter	<0.001
	Classification & Staging	0.98*	2% higher	2% shorter	0.012
	Provisioning & Model Risk	0.98**	2% higher	2% shorter	0.004
	NPE & Collateral Management	0.97***	3% higher	3% shorter	<0.001
	Credit Lifecycle Management	0.97***	3% higher	3% shorter	<0.001
Process originator (reference category: OSI)	Horizontal	1.03***	3% lower	3% longer	<0.001
	JST	1.01*	1% lower	1% longer	0.049
	SREP	1.04***	4% lower	4% longer	<0.001
Measure characteristics					
Severity (reference category: moderate (M2))	Low severity (M1)	0.98***	2% higher	2% shorter	<0.001
	High severity (M3)	1.01*	1% lower	1% longer	0.015
	Very high severity (M4)	1.02†	2% lower	2% longer	0.062
Legal enforceability (reference category: obligation)	Recommendation	1.00	0% lower	0% shorter	0.696
Macroeconomic and market controls					
GDP growth		1.00*	0.1% higher	0.1% shorter	0.029
Unemployment rate		1.00†	0.1% lower	0.1% longer	0.062
Herfindahl index		1.00*	0.4% lower	0.4% longer	0.041
Bank characteristics					
Governance score		0.99***	1.2% higher	1.2% shorter	<0.001
Credit risk materiality		0.99*	0.5% higher	0.5% shorter	0.012
Asset turnover		1.00†	0.3% higher	0.3% shorter	0.055
Year fixed effects (ref. year 2016)					
2017		1.04***	4% lower	4% longer	<0.001
2018		1.08***	8% lower	8% longer	<0.001
2019		1.09***	9% lower	9% longer	<0.001
2020		1.06***	6% lower	6% longer	<0.001

Regressors	Time Ratio	Change in Hazard Rate	Change in Time to Closure	p-value
2021	1.05***	5% lower	5% longer	<0.001
2022	1.05***	5% lower	5% longer	<0.001
2023	1.04***	4% lower	4% longer	<0.001
2024	1.00	0% higher	0% longer	0.795
2025	0.95*	5% higher	5% shorter	0.012
Observations	3,744			
Events	3,744			
Distribution	Weibull AFT			
Log-likelihood ratio test	χ^2 = 396.4 (28 df)			<0.001
Significance	† p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001			

Acknowledgements

We thank Klaus Düllmann, Spyros Palligkinis and Anne Puustelli for comments on a previous version of the paper. We would like to thank all the discussions, comments and contributions from the BCBS Research Group Workstream on Supervisory Effectiveness, the ECB Banking Supervision Internal Research Seminar, the Directorate Supervisory Strategy and Risk (DSSR) National Competent Authorities' network, the DSSR's Extended Management group and the Strategic Risk and Analytics Division. Your inputs have helped to make this paper more thorough and targeted in tackling our intended goal.

Alessandra Agati
European Central Bank, Frankfurt am Main, Germany; email: alessandra.agati@ecb.europa.eu

Adele Fontana
European Central Bank, Frankfurt am Main, Germany; email: adele.fontana@ecb.europa.eu

João Matos Leite
European Central Bank, Frankfurt am Main, Germany; email: joao.leite@ecb.europa.eu

Claudia Schwarz
European Central Bank, Frankfurt am Main, Germany; email: claudia.schwarz@ecb.europa.eu

© European Central Bank, 2026

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

All rights reserved. Any reproduction, publication and reprint in the form of a different publication, whether printed or produced electronically, in whole or in part, is permitted only with the explicit written authorisation of the ECB or the authors.

This paper can be downloaded without charge from www.ecb.europa.eu, from the Social Science Research Network electronic library or from RePEc: Research Papers in Economics. Information on all of the papers published in the ECB Working Paper Series can be found on the ECB's website.

PDF ISBN 978-92-899-8072-2 ISSN 1725-2806 doi:10.2866/7323114 QB-01-26-252-EN-N