READER GUIDE

AUDIENCE	Executive / Practitioner
PRIMARY READER	CROs, consent-order monitors, governance committees, and programme executives new to the research series
ALSO RELEVANT FOR	Any reader who has encountered one of the finding papers and wants context for the full series.
WHAT IT COVERS	The entry point to the research series. No statistical knowledge required. Explains what was found and how to navigate all 22 papers.

WP-0 · ALLAZO MANAGEMENT RESEARCH SERIES

The State of Programme Reporting: An Industry Baseline

Entry point — why the industry has not measured what RAG status actually means

22	6,063	173	3	34	0
WHITE PAPERS IN SERIES	STATUS ENTRIES ANALYSED	PROJECTS TRACKED	YEARS OF OBSERVATION	STATISTICS AUDITED	DISCREPANCIES IN AUDIT

SECTION 01

The Problem This Research Addresses

Programme governance in regulated financial institutions runs on RAG status — the Red, Amber, Green traffic-light system that summarises the health of each project or workstream in a reporting pack. It is the primary information signal that senior executives, risk officers, and regulatory monitors receive about programme delivery. It informs governance decisions, consent-order progress reports, and SteerCo discussions about where to direct attention and resources.

The problem is that no one had systematically measured how reliable that signal is.

The industry has produced substantial guidance on how to define RAG status, what criteria should trigger each colour, and how governance committees should respond to different status combinations. What the industry has *not* done — until this research — is take a large sequential dataset of actual RAG entries and ask: how often does a Green entry turn out to be wrong? How often does a declared recovery fail? How long does a programme typically have between its first warning signal and its first crisis? Does the pattern of reporting predict the pattern of outcomes?

These are empirical questions. They require data to answer. This research programme collects that data — **6,063 RAG status entries from 173 projects across a Tier 1 capital markets division, spanning three years** — and answers them.

SECTION 02

What the Existing Literature Says

The programme management literature is not silent on the question of programme failure. Several significant studies have documented failure rates. The Standish CHAOS reports have tracked IT project outcomes since 1994. McKinsey and Oxford's analysis of large IT programmes found significant overrun rates in their sample. BCG's work on programme management office design identified common failure modes in large-scale transformations. Bent Flyvbjerg's research on megaproject delivery — now extended to software and IT — documents systematic optimism bias in programme forecasting across decades and geographies.

Each of these studies has made an important contribution. But they share a common limitation for the governance question this research addresses: **they measure outcomes, not signals.** They tell you how many programmes overran their budget or missed their deadline. They do not tell you what the programme's RAG reporting showed in the periods *before* it overran — whether the signals were there and missed, or whether the reporting architecture failed to produce signals at all.

STANDISH CHAOS	MCKINSEY / OXFORD	BCG PMO IMPERATIVE	FLYVBJERG MEGAPROJECT
~31% of IT projects delivered successfully (on time, on budget, on scope). Since 1994.	**~45% average cost overrun for large IT programmes.**	**~70% of strategic initiatives fail to achieve expected value.**	**Systematic optimism bias documented across decades and industries.**
Documents the scale of delivery failure. Does not examine whether the reporting system predicted it.	Identifies governance design as a factor — but does not test whether RAG reporting quality explains the variance.	Measures outcomes — not whether governance reporting flagged deterioration.	Measures forecast bias — not reporting reliability once a project is live.

The gap the Allazo research fills is specific: **not what proportion of programmes fail, but what the reporting architecture shows before and during failure.** The existing literature tells us the patient often dies. This research examines whether the diagnosis was correct.

SECTION 03

The Research Programme

The Allazo Programme Intelligence Research Series is based on a single dataset: **6,063 RAG status entries from 173 projects across 245 programmes** in a Tier 1 capital markets division, covering January 2023 through early 2026. The dataset was compiled from programme reporting under institutional confidentiality obligations; the underlying data is not available for third-party distribution. The full methodology — including all numerators, denominators, exclusion criteria, and confidence intervals — is disclosed in every paper.

The research was conducted in three phases. The first phase established five core empirical findings: the Phantom Green rate, the False Recovery rate, the Tipping Point, the Delivery Clock, and the Programme DNA finding. These findings are documented in individual white papers (WP-B through WP-O) and are the primary commercial output of the series.

The second phase extended the analysis using more advanced statistical methods: survival analysis, change-point detection, second-order Markov modelling, and portfolio architecture analysis. These findings are in WP-P through WP-S. The third phase applied a mathematical model — a Continuous-Time Markov Chain — to the data, producing the first quantified estimates of governance intervention effects on long-run portfolio health (WP-T). The series closes with WP-U on generalisation and scope and WP-V on the unified governance defect thesis.

27.8%	36.4%	3.32×	25.5%
PHANTOM GREEN	**FALSE RECOVERY**	**TIPPING POINT**	**NO-WARNING RED**
Green entries contradicted next	Recoveries reverse within 2 periods	Risk multiplier after a 3-period Amber streak	Red events with no prior Amber warning

On reproducibility: Every statistic in the research series was independently recomputed from the source dataset using a documented audit script. All 34 published statistics were tested; **34 were confirmed. Zero discrepancies.** All conclusions confirmed. The full methodology and audit details are available on request.

SECTION 04

The Five Core Findings

The five core findings are the commercial centrepiece of the research. Each one identifies a specific failure mode in how programme health is reported — and each one supports a specific governance protocol designed to address that failure mode.

Phantom Green: 27.8% of Green Entries Are Immediately Contradicted

Nearly three in ten Green RAG entries are followed in the very next reporting period by an Amber or Red entry for the same project. The programme reports health it does not have. This is not a small-scale measurement error — **27.8%** is a consistent, bootstrapped rate across three years and 3,073 Green transitions, with a 95% confidence interval of [22.3%, 25.4%]. The rate rose to 34.7% in 2025. The governance response is the *Evidence-Backed Green* standard: requiring a named, verifiable deliverable alongside each Green declaration rather than accepting managerial assessment alone.

False Recovery: 36.4% of Recovery Declarations Fail Within Two Periods

More than one in three Amber-to-Green transitions reverses to Amber or Red within two reporting periods. A declared recovery is not a confirmed recovery — it is a stated intention that the data shows fails more than one-third of the time. In the 2024–2025 out-of-sample validation period, this rate rose to 52.7%: roughly a coin flip. The governance response is the *Two-Period Hold*: a recovery is reported as "Under Confirmation" until it has held for two consecutive periods with evidence.

The Tipping Point: Amber Streak of Three Is a Red Predictor

Projects with a maximum Amber streak of three or more consecutive reporting periods are **3.32 times** more likely to eventually reach Red than projects without such a streak (80.2% versus 24.7% ever-Red rate). The streak self-reinforces: the Amber-to-Amber persistence rate is 60%, meaning each period in Amber makes the next one more likely. The governance response is mandatory escalation at streak-3: a named accountable owner, a specific resolution date, and a dedicated governance agenda item — regardless of whether the programme manager believes the Amber is temporary.

Delivery Clock: Median 14 Periods from First Amber to First Red

The median time between a project's first Amber entry and its first Red event is **14 reporting periods**. In a weekly cycle, that is fourteen weeks. But **25.5%** of Red events occur with no prior Amber warning at all — the programme reports Green and then Red with nothing in between. For projects that do show warning, the window is real but shorter than most frameworks assume: the 25th percentile is just 5 periods. The governance response is weekly escalation review for any Amber project — monthly SteerCo review is too slow for the lower quartile of the warning window.

Programme DNA: Starting State Predicts Lifetime Trajectory

A programme's starting RAG status is the single strongest predictor of its lifetime health trajectory. Green-start programmes spend 65.4% of their reporting life in Green. Amber-start programmes spend 47.3% — an 18 percentage-point gap that persists across the entire observation window. Every Red-start programme in the dataset reached at least one Red event: **100% ever-Red across all Red-start projects**, with no exceptions. The governance response is treating the launch RAG declaration with the same evidential standard as any in-flight declaration — and resourcing Amber-start programmes with the explicit expectation that they will require more governance throughout their life.

SECTION 05

What Makes This Research Different

Three things distinguish this research from the existing programme management literature.

First, it is sequential. The dataset contains complete reporting histories for each project — every entry, in order, across three years. This allows the research to ask temporal questions that cross-sectional studies cannot: not just "what proportion of projects failed?" but "how long before failure did the signal appear, and what did that signal look like?" Sequential data is necessary for the Delivery Clock, the Tipping Point, the Phantom Green, and the False Recovery findings. None is possible from outcome data alone.

Second, it is reproducible. Every statistic in the series has been independently recomputed from the raw data using a documented audit script. Any party with access to a comparable sequential RAG dataset can apply the identical methodology and test whether the findings replicate. This is a higher standard of transparency than the Standish CHAOS reports (which have not disclosed their underlying data for thirty years) and comparable to academic research standards that most practitioner publications do not meet.

Third, it measures the reporting architecture, not just the outcomes. The existing literature documents that programmes fail. This research documents the pattern of the signals that preceded or failed to precede the failure. That distinction is what makes the findings actionable: knowing that 27.8% of Green entries are Phantom Green tells you specifically what to fix in the governance framework. Knowing that 31% of programmes fail tells you that something went wrong but not what or where.

SECTION 06

What to Read Next

The research series is organised for different readers with different needs. The following map shows the recommended reading paths based on your primary interest.

CORE FINDINGS — READ FIRST	GOVERNANCE PROTOCOLS — WHAT TO DO
WP-D · WP-O · WP-E · WP-B · WP-C	**WP-A · WP-F · WP-H · WP-J · WP-M**
The five empirical findings, each in its own paper. These are the [illegible]	[illegible]
ADVANCED ANALYTICS — DEEPER EVIDENCE	**MATHEMATICAL MODELLING — SPECIALIST READERS**
WP-L · WP-I · WP-N · WP-P · WP-Q · WP-R · WP-S	**WP-T · WP-U · WP-V**
Survival analysis, structural break detection, portfolio architecture, [illegible]	[illegible]

If you are evaluating this research for a consent-order context: Start with WP-D (Phantom Green) and WP-O (False Recovery). These two papers address the most common regulatory concern — that programme health reporting overstates programme health — with specific, quantified evidence and specific governance protocols. Follow with WP-B (Delivery Clock) for the governance timing implications, and WP-A for the integrated framework. The Methodology and Reproducibility Statement, available on request, provides the full audit trail for any examiner or monitor who wishes to verify the underlying statistics.

SECTION 07

A Note on Confidentiality and Reproducibility

The dataset underlying this research was compiled from programme reporting within a Tier 1 capital markets division under institutional confidentiality obligations that survive the engagement period. The institution cannot be named. The underlying data is not available for third-party distribution.

This confidentiality is consistent with how research of this type is typically conducted. The Standish Group has maintained the confidentiality of its underlying project data for over thirty years while publishing research derived from it. Academic clinical researchers routinely publish findings from datasets that are not publicly available, subject to IRB oversight and methodology disclosure requirements.

What is available — and what the Standish reports typically do not provide — is the **complete methodology**: every numerator, denominator, exclusion criterion, and confidence interval used to produce every published statistic. The reproducibility audit script, which recomputes all 34 published statistics from the raw data and produces a pass/fail comparison table, is available on request to any party with an appropriate confidentiality agreement. Any organisation with access to a comparable sequential RAG dataset can apply the identical methodology and test whether the findings replicate in their context.

The data is confidential. The methodology is not. That is the appropriate balance for research of this type, and it is a stronger position than most practitioner research takes.