

The Interpretive Limits of Expected Goals: Model Dependence, Realization Variability, and Decision-Making Under Uncertainty

giovanni gallo

GIOVANNI.GALL0275@GMAIL.COM

Quantum Sport Analysis <https://orcid.org/0009-0009-8689-8563>

Research Article

Keywords: expected goals, xG, football analytics, aleatoric uncertainty, betting, probabilistic models, realization divergence, human judgment, live betting, decision-making

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11153008/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

The Interpretive Limits of Expected Goals: Model Dependence, Realization Variability, and Decision-Making Under Uncertainty

Implications for Post-Match Analysis and Live Football Betting

Preprint v1.0 — 24 September 2026

Author: Giovanni Gallo

Affiliation: Quantum Sport Analysis

ORCID: 0009-0009-8689-8563

Abstract

Expected Goals (xG) models provide conditional probability estimates of scoring based on a finite modelled information set. Contemporary implementations can incorporate rich contextual, defensive, goalkeeper, sequential and tracking information; this is a genuine methodological advance, but it does not remove the conditionality of the resulting estimate with respect to the represented information, model specification and reference data. Their outputs are therefore probabilistic, model-dependent representations rather than deterministic statements about what an individual football event or match should produce. This distinction is relevant both to post-match statistical interpretation and to live betting, where a precise numerical display can become a decision cue.

This paper combines a methodological framework with a preliminary retrospective screening of 154 xG-complete matches from the first four rounds of the 2026/27 Premier League, Serie A, La Liga and Bundesliga, yielding 308 team-match observations. Cumulative xG was mapped to integer goal regions separated by half-goal boundaries (0.5, 1.5, 2.5, 3.5, ...). These cutpoints were selected because half-goal lines create unambiguous win/loss settlement boundaries in football goal markets, including team-total markets that correspond directly to the team-match unit analysed here. The transformation is analytical; it is neither a bookmaker pricing model nor a native output of the underlying xG models. Realized-goal divergence was defined as $\Delta_G = G_{\text{actual}} - B(xG)$.

Exact correspondence with the half-goal-threshold xG analytical band occurred in 35.7% of observations; 83.4% were within one goal and 16.6% were at least two goals away. Mean absolute band divergence was 0.834 goals, while signed mean divergence was only +0.055, showing that aggregate cancellation can coexist with substantial individual realization dispersion. A didactic Poisson reference with $\lambda = xG$ produces an expected exact-band rate of approximately 34.2%. This reference is used only to show that a low exact-realization rate can arise naturally in a probabilistic count process; it is not a calibration test or generative model for the provider xG values. Sensitivity analyses excluding observations close to half-goal cutpoints produced materially similar results. A secondary two-sided threshold screen showed material variation by both line and side, with marked class imbalance at the higher thresholds.

The second part proposes a stratified prospective experiment crossing three participant strata—football experts, regular bettors/football fans and novices—with three randomized information environments within each stratum: match visible without analytical statistics, match visible with statistics overlaid, and statistics-only from bookmaker/provider sources. Three distinct decision tasks are studied: final-match Over/Under 2.5 at 15 minutes, full-time 1X2 at half-time, and next-goal prediction at 60 minutes. A separate out-of-sample modelling benchmark estimates the incremental predictive contribution of xG to each target. The central claim is not that xG is

inaccurate or simplistic, but that even a sophisticated xG estimate remains a conditional statistical representation rather than an observed realized outcome or a direct probability of a future betting target. The decision impact of statistical information must therefore be evaluated within the environment in which it is used.

Keywords: expected goals; xG; football analytics; aleatoric uncertainty; betting; probabilistic models; realization divergence; human judgment; live betting; decision-making

1. Introduction

Expected Goals (xG) have become a standard component of football analytics. At shot level, an xG model estimates a probability of scoring conditional on a selected representation of the action. In simplified form,

$$\hat{p}_i = \hat{P}(G_i = 1 \mid X_i),$$

where G_i is the binary goal outcome and X_i is the information made available to the model. Depending on the modelling system, X_i may include location, angle, body part, preceding action, defensive configuration, goalkeeper information, or other contextual variables (Anzer & Bauer, 2021; Mead et al., 2023; Cefis & Carpita, 2025; Lago-Peñas et al., 2026). Contemporary xG modelling spans a continuum from relatively sparse event-based representations to context-enhanced, event-sequence and tracking-based systems that can encode substantially richer spatial and temporal state information (Anzer & Bauer, 2021; Bandara et al., 2024). The argument developed here therefore does not depend on xG being a simplistic metric.

The analytical value of xG is clear. A final score contains very little information about the quality and frequency of the chances that preceded it, whereas xG provides a probabilistic summary of shot opportunities and can support performance evaluation under outcome randomness (Brechot & Flepp, 2020). The interpretive problem begins when an expected value is treated as though it were a deterministic statement about a single realization.

This problem appears in at least two contexts.

First, **post-match statistical analysis** frequently places cumulative xG beside the final score. These are different statistical objects: the final score is the observed realized outcome of the completed match, whereas xG is a conditional statistical representation derived from a finite information set. A large numerical difference between them does not make the comparison meaningless, but it does prevent the xG value from being read as the score that deterministically “should” have occurred.

Second, **live betting** places statistical representations inside a real-time decision environment. A bettor acts on one currently unfolding match, under changing odds and time pressure. The prospective experiment therefore compares decisions made from direct match observation, direct observation plus statistical overlays, and statistics-only representations supplied by bookmaker/provider interfaces.

The present paper treats these as successive information transformations:

football event state $\rightarrow$ statistical representation $\rightarrow$ aleatoric realization

and, when the statistic is displayed,

statistical representation $\rightarrow$ human judgment $\rightarrow$ betting decision.

Figure 1 summarizes this architecture.

Figure 1. Conceptual architecture of xG representation and use

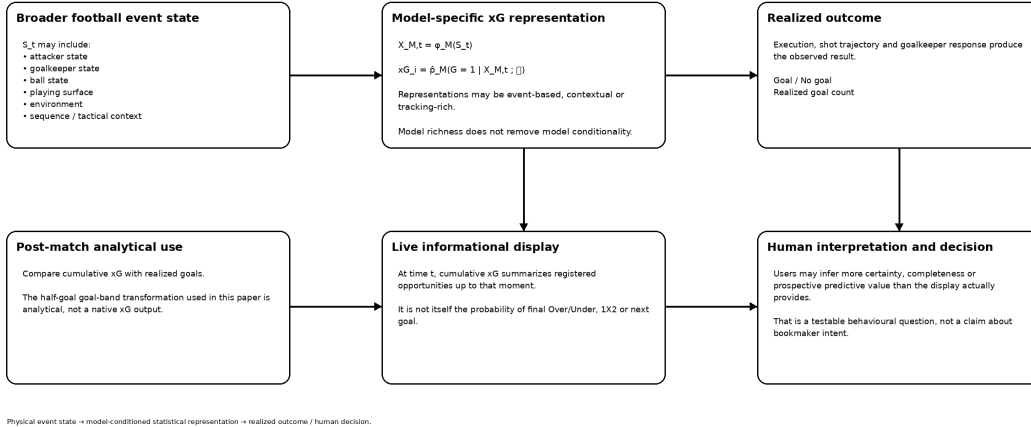


Figure 1. Conceptual architecture of the paper. The upper row separates the broader football event state, the model-specific xG representation and the realized outcome. The lower row distinguishes post-match analytical use from live display and subsequent human interpretation. Modern xG representations may be information-rich, but the resulting estimate remains conditional on the represented state, model specification and reference data. The retrospective half-goal transformation is analytical and is not a native xG output.

A more explicit model-level notation is:

$$\hat{p}_{i,M,\mathcal{D}} = \hat{P}_M(G_i = 1 | X_i; \mathcal{D}),$$

where M denotes the model specification and $\mathcal{D}$ the training or reference data. This notation separates the model output from an idealized event-specific probability. The number displayed to a user is therefore inseparable from the information set, model architecture and reference population used to produce it.

The empirical component of the study does **not** test shot-level xG calibration. Instead, it examines what happens when cumulative team xG is translated into discrete goal regions separated by half-goal cutpoints. The use of 0.5, 1.5, 2.5, 3.5 and higher half-goal thresholds is deliberate: these numerical boundaries correspond to common settlement boundaries in football goal markets and, at the team-match level used here, most directly to team-total goal lines. The purpose is not to claim that bookmakers convert xG into these bands, but to study how a continuous probabilistic quantity behaves when projected onto discrete, betting-relevant goal regions.

The first research question is therefore:

When cumulative xG is mapped onto half-goal-threshold analytical bands, what is the distribution of realized team goals around those bands?

The second is:

To what extent does aggregate signed balance conceal the distribution and magnitude of team-match realization divergence in the present sample?

The third concerns human use:

How do probabilistic judgments and betting decisions differ across match-only, match-plus-statistics and statistics-only information environments, and how do those differences vary across participant expertise strata?

The contribution is not a new xG estimator. It is the integration of three analytically distinct layers: (1) model and information-set dependence; (2) quantitative illustration of team-match realization dispersion around half-goal analytical boundaries; and (3) a prospective framework comparing football decisions across three information environments in which statistical representations, including xG, are absent, overlaid on the match, or presented without match video. This differs from treating xG either purely as a modelling problem or purely as a gambling cue.

The remainder of the paper develops these three layers in sequence.

2. Expected Goals as Conditional Statistical Representations

2.1. Theoretical conditional probability and estimated xG

It is useful to distinguish the theoretical conditional probability from the output of a fitted model.

Let

$$p^*(X) = P(G = 1 \mid X)$$

denote the population conditional probability under the specified information set X . If the broader event state is represented as

$$S = (X, Z),$$

where Z contains relevant state information not included in X , then the law of total probability gives:

$$p^*(X) = P(G = 1 \mid X) = \int P(G = 1 \mid X, Z)P(Z \mid X) dZ.$$

A fitted xG model does not automatically equal $p^*(X)$. Its output is better written as:

$$\hat{p}_M(X; \mathcal{D}),$$

which is an estimate produced by model M from reference data $\mathcal{D}$.

This distinction matters because two sources of uncertainty coexist. Outcome variability remains even if $p^*(X)$ were known exactly, while the estimate $\hat{p}_M$ can additionally differ from $p^*(X)$ because of model specification, finite data, measurement, annotation and calibration.

2.2. Model richness, conditionality and provider dependence

A numerical xG value is therefore conditional on more than the visible football action:

$$xG_i = \hat{p}_M(X_i; \mathcal{D}).$$

Different providers can legitimately return different values for the same real-world event if they differ in:

- operational event definitions;
- available features;
- measurement resolution;
- treatment of defensive or goalkeeper context;
- sequence information;
- training data;
- modelling algorithm;
- post-processing or calibration.

Provider disagreement is therefore conceptually expected. Numerical precision does not imply model uniqueness.

Contemporary xG systems can nevertheless be highly sophisticated. Published approaches use synchronized positional and event data, defensive-pressure information, goalkeeper positioning, preceding-event context, event sequences and tracking-derived spatial information (Anzer & Bauer, 2021; Bandara et al., 2024). This represents real methodological progress because a richer state representation can reduce the amount of relevant information that remains outside the model.

For conceptual purposes, progressively richer representations can be written as:

$$X_1 \subset X_2 \subset X_3 \subset S,$$

where X_1 denotes a basic event representation, X_2 a richer contextual representation, X_3 an advanced dynamic or tracking representation, and S the broader event state. This nesting is an expository construction rather than a claim that the feature sets of real commercial providers are literally nested.

The key methodological point is:

$$\text{model richness} \neq \text{model independence}.$$

Or, more directly:

Model richness does not remove model conditionality.

Even a richly specified estimate remains conditional on the state represented in X , on the fitted model M , and on the reference or training data $\mathcal{D}$.

A related point concerns player- and goalkeeper-specific heterogeneity. Let A denote attacker-specific state or capacity and K goalkeeper-specific state or capacity. If these quantities are not included in the conditioning set, then the theoretical conditional probability can be written as:

$$p^*(X) = \iint P(G = 1 \mid X, A, K) P(A, K \mid X) dA dK.$$

Thus:

$$P(G = 1 \mid X) \neq P(G = 1 \mid X, \text{this specific attacker, this specific goalkeeper})$$

unless the relevant attacker- and goalkeeper-specific information is explicitly conditioned upon. This is a statement about conditional probability, not a claim that every fitted xG model performs this marginalization perfectly. The fitted output $\hat{p}_M(X; \mathcal{D})$ can still differ from $p^*(X)$.

The expression “average player” is therefore too crude as a universal description. A more precise statement is that, when player-specific execution capacity or identity is excluded from the conditioning set, player heterogeneity is absorbed into the reference-population relationship represented by the model. Scholtes and Karakuş (2024), for example, report player-specific effects that remain detectable after richer shot-related covariates are introduced. Conversely, some provider models deliberately exclude shooter identity: the public xGStat documentation states that the same model-represented chance receives the same xG value regardless of the player taking the shot (xGStat, 2026).

2.3. Population calibration is not individual realization certainty

Probabilistic calibration is a repeated-event property. If events assigned probability 0.20 occur approximately 20% of the time in the relevant reference population, the model may be calibrated even though four out of five individual events do not become goals.

Accordingly,

$$\text{population calibration} \neq \text{certainty of an individual realization.}$$

This distinction survives aggregation. Cumulative xG can be written as:

$$XG_t = \sum_{i=1}^{n_t} \hat{p}_i.$$

Conditional on the fitted model outputs, this sum is the model-implied expected goal count for the registered shot sequence. Linearity of expectation does not require the shot outcomes to be independent. Independence becomes relevant only when constructing a full count distribution from the constituent shot probabilities. The cumulative value is not:

$$P(G_{\text{team}} = XG_t),$$

nor is it automatically a win probability, next-goal probability, or final-total probability.

For example,

$$1.00 = 2 \times 0.50$$

and

$$1.00 = 10 \times 0.10$$

have the same cumulative expected value but very different internal opportunity structures.

2.4. Post-match interpretation

The distinction becomes important when cumulative xG is used retrospectively. Statements such as “the team should have scored two goals because it produced 2.0 xG” convert an expected value into a single-outcome claim.

The correct probabilistic interpretation is weaker: the observed shot sequence generated a collection of model-estimated scoring probabilities whose sum equals approximately 2.0. Multiple realized goal counts remain possible.

Therefore:

$$\text{expected value} \neq \text{most certain individual realization.}$$

The empirical screening below quantifies this point after projecting cumulative xG onto discrete half-goal-derived goal regions.

2.5. Representational and annotation uncertainty

The model input itself may already be a representation rather than a raw physical measurement:

$$\text{physical event} \rightarrow \text{measurement} / \text{annotation} \rightarrow X \rightarrow \hat{p}_M.$$

A directly measured geometric variable is conceptually different from an operational category assigned through an annotation procedure. This does not make categorical variables invalid. It means that uncertainty can arise before statistical estimation and that provider dependence can partly reflect differences in event representation.

No inter-annotator or inter-provider reliability estimate is attempted in the present study.

2.6. Aleatoric and epistemic uncertainty

The central distinction for this paper is between **aleatoric** and **epistemic/model** uncertainty (Hüllermeier & Waegeman, 2021).

Aleatoric uncertainty refers to variation in realized outcomes even when the conditioning state and probability are fixed. A shot with true conditional scoring probability $p = 0.20$ can become a goal or not become a goal; neither individual outcome contradicts the probability statement.

Epistemic uncertainty concerns uncertainty about the probability estimate itself: the information set, model form, training population, measurement process, annotation scheme, parameter estimation and calibration.

Thus:

$$\text{aleatoric realization uncertainty} \neq \text{uncertainty about } \hat{p}_M.$$

The retrospective analysis characterizes realized dispersion that is consistent with aleatoric variation, but the design does not decompose observed divergence into aleatoric and epistemic components.

2.7. Numerical precision is not certainty

A point estimate such as 1.73 xG communicates numerical precision but does not reveal:

- outcome variance;
- model uncertainty;
- provider disagreement;
- reference-population uncertainty;
- structural uncertainty in the information set.

Therefore:

$$\text{decimal precision} \neq \text{predictive certainty}.$$

This becomes especially important once the value is presented as a live decision cue.

3. From Model-Represented State to the Broader Physical Event State

3.1. Model-represented state and broader event state

The relevant methodological distinction is not that xG models are static by definition. Contemporary xG modelling ranges from sparse event-based representations to systems that incorporate richer contextual, sequential and tracking information (Anzer & Bauer, 2021; Bandara et al., 2024).

A more general representation is:

$$X_{M,t} = \phi_M(S_t),$$

where S_t denotes the broader physical and behavioural state of the football event at time t , and $X_{M,t}$ is the representation of that state made available to model M . The mapping ϕ_M can be very rich, but it remains model-specific.

Conceptually, the broader state may contain attacker state, goalkeeper state, tactical and sequence context, playing-surface state, environmental conditions, and ball/equipment state. These components need not be mutually exclusive, and an advanced model may already encode part of them directly or indirectly.

The methodological question is therefore not whether modern xG is “only distance and angle.” It is whether the model-specific representation $X_{M,t}$ exhausts the state that is relevant to the event-specific probability of scoring.

3.2. Player- and goalkeeper-specific execution state

A shot is the endpoint of a dynamic execution process involving movement, balance, orientation, timing, support-foot mechanics and interaction with opponents. Football kicking performance depends on coordinated movement, segmental velocities, support-foot placement and foot-ball impact mechanics (Lees et al., 2010).

Two players can therefore occupy the same externally observed geometry while having different feasible spaces of biomechanical execution. Relevant player-specific differences can include finishing ability, weak-foot competence, first-touch quality, motor consistency, body orientation, support-foot state, movement history and physiological or cognitive condition. Fatigue has been associated with changes in football-specific physical, technical and decision performance, although the magnitude and direction of those effects depend on the task and study design (Dambroz et al., 2022; Smith et al., 2016; Peng et al., 2026).

The same logic applies to the goalkeeper. Goalkeeper position is an important predictor and is already incorporated in some advanced xG models (Anzer & Bauer, 2021), but position is only one component of goalkeeper state. Reach, preparatory posture, previous movement, momentum, orientation, reaction state and the mechanics of the subsequent dive can differ even from similar nominal locations. Biomechanical studies of goalkeeper diving show that save performance depends on coordinated lower-limb power, angular velocity, preparatory posture and movement direction (Ibrahim et al., 2020).

These observations do not imply that all such variables are absent from all xG systems. They establish a conditionality point: the event-specific probability represented by a model depends on which components of player and goalkeeper state are actually encoded in that model’s information set.

3.3. Ball, playing surface and environmental state

A football event is also a physical interaction among the player, ball, playing surface and surrounding environment.

Ball construction and surface texture influence aerodynamic forces and flight trajectory. Wind-tunnel evidence shows that changing soccer-ball surface texture can materially alter aerodynamic behaviour and calculated trajectories (Hong et al., 2019). Ball inflation pressure is regulated but not fixed to one universal exact value: IFAB Law 2 permits a range of 0.6-1.1 atmosphere at sea level (IFAB, n.d.). Experimental work also shows that inflation pressure can alter impact mechanics in heading contexts (Pereira et al., 2024). These findings establish physical relevance; they do not provide an xG effect size for ordinary foot-shot conversion.

Playing-surface state is likewise not reducible to a binary stadium identifier. Natural, reinforced,

hybrid and synthetic surfaces can differ in hardness, stability, evenness, moisture, traction and ball-surface interaction. FIFA’s testing programmes explicitly measure surface hardness, stability, evenness and ball-surface interaction for natural surfaces, and player-surface and ball-surface interaction for football turf (FIFA, 2021; FIFA, n.d.). A systematic review of cleat-surface interaction reports performance differences associated with traction, surface type and wet conditions, including effects on ball velocity and kicking accuracy in the underlying literature (Silva et al., 2017).

Environmental conditions such as temperature, humidity, precipitation and altitude can also affect physical performance and player pacing; effects on technical performance are more heterogeneous across studies (Illmer & Daumann, 2022).

The conservative physical statement is therefore:

Identical observed shot geometry does not imply an identical player-goalskeeper-ball-surface-environment physical state.

Because scoring probability depends on execution and ball trajectory, differences in that broader physical state are relevant to the event-specific probability of scoring. The present study does not estimate the marginal effect of any of these variables on goal probability.

A second distinction is essential:

physical relevance $\neq$ documented model inclusion.

A factor may be physically relevant to shot execution while being directly represented, indirectly proxied, absent from, or simply undocumented in a particular xG model. Whether the numerical xG reported by a provider changes with surface, weather or ball state is therefore a separate model-specific question.

3.4. Chance formation, execution and post-shot trajectory

The scoring process can be separated conceptually into:

chance formation $\rightarrow$ execution $\rightarrow$ post-shot trajectory.

Pre-shot xG and post-shot probability answer different questions because post-shot models can condition on information that did not exist before ball contact. A pre-shot model should therefore not be evaluated as though it had access to the realized trajectory.

The broader physical-state discussion above does not turn these variables into post-hoc explanations for any individual miss or goal. It establishes that the model-represented state and the full physical realization need not be identical objects.

3.5. Observability boundary of shot-based xG

Shot-based xG is bounded by its unit of observation. Consider two otherwise similar attacking transitions. In the first, the final pass reaches the receiver and a shot is recorded. In the second, the pass is misplaced by a small distance and the receiver cannot shoot. The second sequence produces no registered shot and therefore no conventional shot xG for that terminal action.

Thus:

dangerous physical state $\neq$ registered shot event.

This is an observability boundary of shot-based xG, not a failure of football analytics as a whole. Possession-value, expected-threat and related frameworks are specifically designed to assign value to earlier states and actions in attacking sequences. For example, expected-possession-value models can evaluate passes, ball drives and shots using spatiotemporal state information before a terminal shot occurs (Fernández et al., 2021).

3.6. Non-attribution of the empirical divergence

The variables discussed in this section should not be confused with the empirical Δ_G screening reported below.

The retrospective screening does not attribute any observed realization divergence to a specific latent physical, physiological, player-specific, goalkeeper-specific, ball, surface or environmental variable.

The empirical analysis measures realization dispersion after a defined transformation of cumulative xG. It does not identify the causal source of individual deviations, nor does it estimate how much a richer information set would change any provider-reported xG value.

4. Methods

4.1. Design and unit of analysis

A retrospective descriptive screening was conducted at the **team-match** level. The analytical unit is one team’s cumulative xG and realized goals within one match.

The analysis is not a shot-level calibration test.

4.2. Sample

The sample contains the first four official rounds of the 2026/27:

- Premier League;
- Serie A;
- La Liga;
- Bundesliga.

The four-round window was fixed as an early-season exploratory screening frame. It was not selected through outcome-based tuning, is not claimed to represent the full season, and was not determined by an inferential sample-size calculation.

A total of 156 fixtures were recorded. Two Bundesliga fixtures lacked usable xG in the prespecified source and were retained as **SOURCE GAP** without imputation from another provider.

The xG-complete sample therefore contains:

154 matches

and:

308 team-match observations.

League-specific observation counts were:

$$N_{\text{Premier}} = 80, \quad N_{\text{SerieA}} = 80, \quad N_{\text{LaLiga}} = 80, \quad N_{\text{Bundesliga}} = 68.$$

4.3. xG provenance

A single primary source was maintained within each league:

- Premier League: xGStat / Wyscout event data;
- Serie A: MARKSTATS;
- La Liga: MARKSTATS;
- Bundesliga: FotMob advanced statistics / Opta.

Because provider and league vary together, cross-league differences are descriptive and cannot be decomposed into independent league and provider effects.

A targeted public-documentation audit was also conducted to separate **documented model inputs** from **physically relevant variables whose inclusion is not established**. This is not a reverse-engineering exercise and should not be read as a complete description of proprietary model internals.

Source used in the screening	Publicly documented inputs relevant to this paper	Public-documentation status for player identity / ball / surface / weather
xGStat / Premier League source	Distance, angle, body part, shot type, assist type, defensive pressure and goalkeeper positioning	Shooter identity is explicitly stated not to be used. Ball state, playing-surface state and weather are not listed among the six documented inputs (xGStat, 2026).
MARKSTATS / Serie A and La Liga source	A sufficiently detailed public feature specification was not identified in the documentation reviewed for this paper	Implementation of player-specific, ball, surface and weather variables is therefore not publicly verified here .
Opta via FotMob / Bundesliga source	Public Opta documentation describes more than 20 variables, including distance, angle, goalkeeper position, goalmouth visibility, defensive pressure, shot type, pattern of play and the previous action	Ball state, playing-surface state and weather are not listed in the reviewed public explainer; this absence from the public list is not proof that no proprietary implementation uses related information (Opta Analyst, n.d.).

This audit supports a narrower claim than “modern xG ignores context.” Some systems explicitly model rich defensive, goalkeeper and sequential information. The unresolved question for any additional physical-state variable is model-specific: whether it is directly represented, proxied by correlated inputs, absorbed through the reference data, or omitted.

4.4. Half-goal-threshold analytical transformation

Cumulative xG was projected onto integer goal regions separated by half-goal boundaries:

$$0.00 \leq xG \leq 0.50 \Rightarrow B(xG) = 0,$$

$$0.50 < xG \leq 1.50 \Rightarrow B(xG) = 1,$$

$$1.50 < xG \leq 2.50 \Rightarrow B(xG) = 2,$$

$$2.50 < xG \leq 3.50 \Rightarrow B(xG) = 3,$$

with the same structure continued for larger values.

In compact form:

$$B(xG) = \lceil xG - 0.5 \rceil,$$

for $xG \geq 0$, using the interval convention above.

The half-goal cutpoints were selected because lines such as 0.5, 1.5, 2.5 and 3.5 create unambiguous win/loss settlement boundaries in football goal markets (Betfair, 2026). Because the empirical unit is a **team-match observation**, the closest betting analogue is the team-total goal market. The numerical cutpoints are therefore betting-motivated analytical boundaries, not bookmaker-implied probabilities and not a reconstruction of bookmaker pricing.

The transformation is also not a native xG output. Its purpose is to examine how a continuous expected-value statistic maps onto discrete goal regions whose boundaries have a direct operational interpretation in football betting.

Values exactly equal to .50 were retained in the lower region, because crossing the corresponding half-goal boundary requires a strictly larger value.

4.5. Realization divergence

For each team-match observation:

$$\Delta_G = G_{\text{actual}} - B(xG).$$

Thus:

- $\Delta_G = 0$: realized goals coincide with the analytical band;

- $\Delta_G < 0$: realized goals fall below the band;
- $\Delta_G > 0$: realized goals exceed the band.

The complete signed magnitude was retained.

To avoid describing realization difference as a model “error”, the principal absolute summary is termed **Mean Absolute Band Divergence (MABD)**:

$$MABD = \frac{1}{N} \sum_{i=1}^N |\Delta_{G,i}|.$$

Signed mean divergence is:

$$\overline{\Delta_G} = \frac{1}{N} \sum_{i=1}^N \Delta_{G,i}.$$

4.6. Raw-data policy

The primary analysis uses all xG-complete observations.

No observation was removed on the basis of:

- realized goals;
- xG magnitude;
- $|\Delta_G|$;
- distributional extremity;
- whether the observation improved or worsened the reported correspondence.

No trimming, winsorization, outlier exclusion or data-driven threshold optimization was applied.

The only missing observations are the two prespecified Bundesliga source gaps.

4.7. Threshold sensitivity

Because half-goal discretization necessarily introduces threshold discontinuities, a secondary sensitivity analysis was conducted.

The primary analysis was repeated after excluding observations whose xG lay within:

$$\pm 0.05$$

and:

$$\pm 0.10$$

of any half-goal cutpoint.

These exclusions are **not** part of the primary analysis and are not used to tune the thresholds. They test whether the principal descriptive pattern is driven disproportionately by observations located immediately next to a classification boundary.

4.8. Didactic Poisson reference

To place exact-band correspondence in probabilistic context, an illustrative count reference was calculated by assuming:

$$G_i \sim \text{Poisson}(\lambda_i = xG_i).$$

For each observation, the probability of realizing the assigned band $B(xG_i)$, the probability of remaining within one goal of that band, the probability of lying at least two goals away, and expected absolute band divergence were calculated under this reference distribution and then averaged across the 308 observations.

This Poisson construction is **not** treated as the generative model of the observed football goals and is not a validation model for the providers. Cumulative xG is a sum of shot probabilities, and a shot-level Poisson-binomial or other dependence-aware formulation would be more closely aligned with the underlying event structure if all shot-level inputs were available. The Poisson reference is used only as a didactic comparison showing that a relatively low exact-realization rate can arise naturally from probabilistic count dispersion.

4.9. Secondary two-sided half-goal-threshold screening

A secondary descriptive screen used the same half-goal boundaries in binary form.

For each team and each line $L \in \{0.5, 1.5, 2.5, 3.5, 4.5\}$, the xG observation was classified as:

$$xG > L$$

for the Over side, and:

$$xG \leq L$$

for the Under side.

The realized goal count was classified against the same half-goal boundary. Because realized goals are integers, no realized observation can equal a half-goal line.

Two conditional realization rates were then calculated:

- **Over-side realization:** among observations with $xG > L$, the share with realized goals also above L ;
- **Under-side realization:** among observations with $xG \leq L$, the share with realized goals below L .

A pooled same-side agreement rate was also reported as a descriptive summary. Because this raw agreement can be dominated by the more prevalent side of a threshold, a **balanced side realization rate** was also calculated as the arithmetic mean of the Over-side and Under-side realization rates, giving equal weight to both sides of the line.

This is a threshold-side realization screen, not a calibration test, classification model or betting strategy. It contains no market odds, bookmaker margin, line-availability timing, stake optimization

or execution rule. At high goal lines the Under side is also much more prevalent than the Over side, so the overall same-side rate must not be interpreted without the separate conditional rates.

4.10. League comparison

League-specific Δ_G distributions were normalized within league because Bundesliga contributes fewer complete observations.

No inferential ranking of leagues is performed.

4.11. Inferential boundary

The screening does not estimate:

- shot-level calibration;
- Brier score of the provider models;
- log loss of the provider models;
- causal effects of latent physical states;
- win probability;
- next-goal probability;
- betting profitability.

Accordingly:

$$\Delta_G \neq \text{xG calibration error.}$$

It is a realization difference produced after an analytical, half-goal-threshold discretization of cumulative xG.

5. Results

5.1. Realized goals within half-goal-threshold xG analytical bands

Figure 2 shows the realized goal distribution within the four most populated analytical bands.

Figure 2. Observed goals within half-goal-threshold xG analytical bands. Each column contains team-match observations whose cumulative xG fell within the corresponding half-goal region. Stacked segments show realized goals. The cutpoints are analytical boundaries selected because they coincide numerically with common half-goal settlement boundaries in football goal-total markets; they are not deterministic predictions produced by the xG providers.

Band sizes were:

$$N_0 = 29, \quad N_1 = 146, \quad N_2 = 87, \quad N_3 = 34.$$

Band 0 produced zero goals in 69.0% of observations, while 31.0% produced at least one goal.

Band 1 produced exactly one goal in 36.3% of observations. Realizations ranged from zero to five goals.

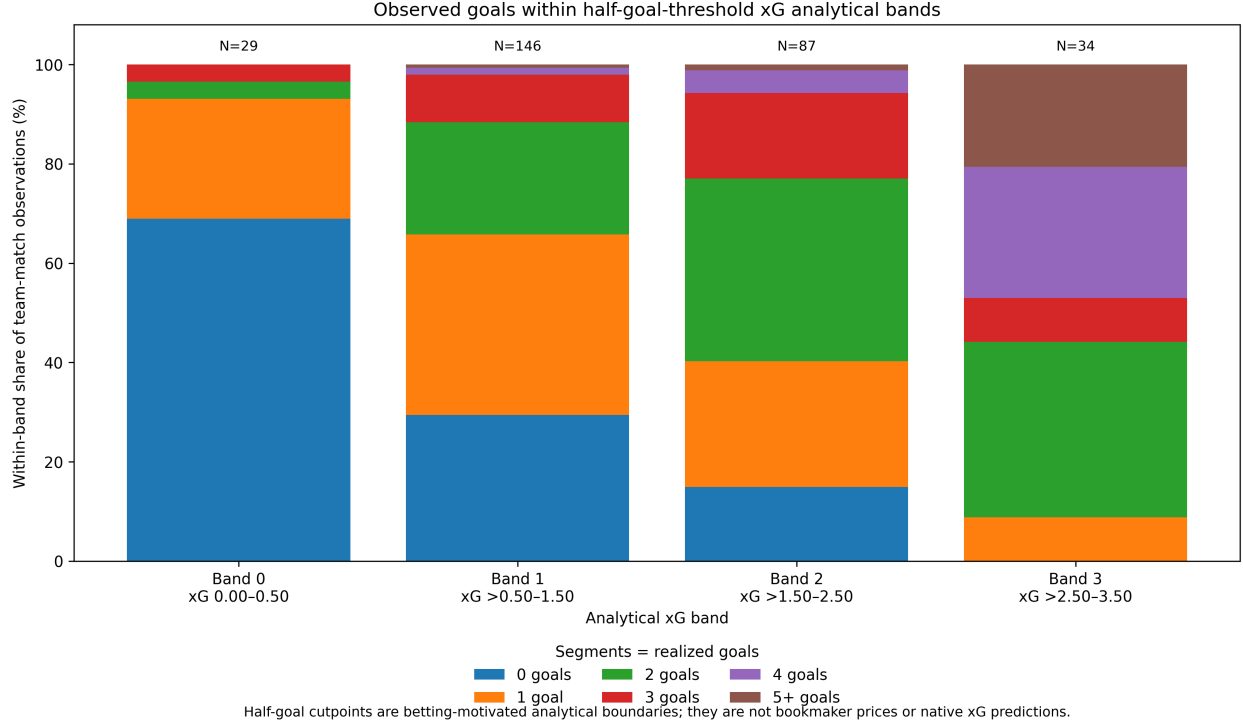


Figure 1: Figure 2. Observed goals within half-goal-threshold xG analytical bands.

Band 2 produced exactly two goals in 36.8% of observations. Its exact rate was therefore almost identical to Band 1 despite a different signed aggregate realization.

Band 3 produced exactly three goals in only 8.8% of its 34 observations, with realized counts distributed on both sides of the analytical band.

5.2. Overall signed realization divergence

Across all 308 team-match observations:

Δ_G	Interpretation	N	%
-3	3 goals below band	1	0.3
-2	2 goals below band	18	5.8
-1	1 goal below band	81	26.3
0	Exact band	110	35.7
+1	1 goal above band	66	21.4
+2	2 goals above band	26	8.4
+3	3 goals above band	5	1.6
+4	4 goals above band	1	0.3

Exact correspondence occurred in:

$$110/308 = 35.7\%.$$

However:

$$257/308 = 83.4\%$$

were within one goal of the assigned band, while:

$$51/308 = 16.6\%$$

were at least two goals away.

The observed mean absolute band divergence was:

$$MABD = 0.834.$$

Signed mean divergence was:

$$\overline{\Delta_G} = +0.055.$$

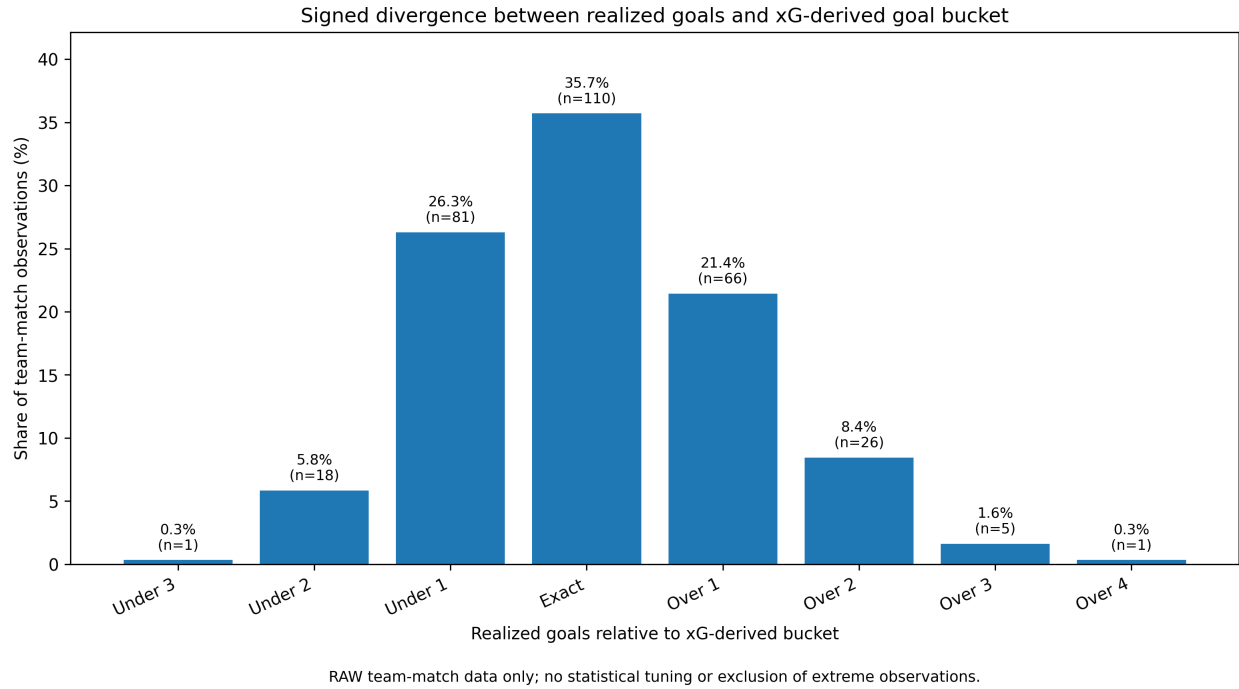


Figure 2: Figure 3. Realized-goal divergence from the half-goal-threshold xG analytical band.

Figure 3. Realized-goal divergence from the half-goal-threshold xG analytical band. Negative values indicate realized goal totals below the assigned analytical band, zero denotes exact correspondence, and positive values indicate realized totals above it. The figure reports the complete raw xG-available sample.

The important combination is therefore:

$$\overline{\Delta_G} \approx 0$$

while:

$$P(\Delta_G = 0) = 0.357.$$

Near-zero aggregate divergence does not imply that individual realizations are concentrated on the assigned band.

5.3. Didactic Poisson reference and probabilistic count dispersion

Under the illustrative assumption:

$$G_i \sim \text{Poisson}(\lambda_i = xG_i),$$

the average probability of realizing the assigned band is approximately:

$$34.2\%.$$

The observed rate was:

$$35.7\%.$$

The corresponding comparison is:

Metric	Observed	Didactic Poisson reference
Exact analytical band	35.7%	34.2%
Within ± 1 goal	83.4%	81.2%
$ \Delta_G \geq 2$	16.6%	18.8%
Mean absolute band divergence	0.834	0.916

The reference is intentionally simple and should not be interpreted as a provider-calibration benchmark. Its role is conceptual: an exact-band rate near one third is not inherently evidence of model failure. It is compatible with substantial aleatoric dispersion around an expected count.

This distinction is central to post-match interpretation. The same xG value can be statistically meaningful even though the single realized score differs from the nearest betting-motivated goal region.

5.4. Threshold sensitivity

Excluding observations immediately adjacent to half-goal cutpoints did not materially alter the principal descriptive results.

Exclusion around each half-goal cutpoint	N retained	Exact band	Within ± 1	$ \Delta_G \geq 2$	MABD	Signed mean
None	308	35.7%	83.4%	16.6%	0.834	+0.055
> 0.05 away	270	34.8%	83.7%	16.3%	0.837	+0.022
> 0.10 away	244	35.7%	84.0%	16.0%	0.824	+0.004

The stability of these values indicates that the aggregate pattern is not driven primarily by xG observations located immediately next to the selected half-goal boundaries.

5.5. Secondary two-sided half-goal-threshold screening

The frozen team-match dataset was also classified on both sides of each half-goal boundary.

Half- goal line	xG Over N	Over realized hit N	Over- side rate	xG Under N	Under realized hit N	Under- side rate	Balanced side rate	Raw same- side rate
0.5	279	223	79.9%	29	20	69.0%	74.4%	78.9%
1.5	133	95	71.4%	175	123	70.3%	70.9%	70.8%
2.5	46	29	63.0%	262	224	85.5%	74.3%	82.1%
3.5	12	6	50.0%	296	272	91.9%	70.9%	90.3%
4.5	3	1	33.3%	305	293	96.1%	64.7%	95.5%

At 0.5, 1.5 and 2.5, the Over-side realization rates were 79.9%, 71.4% and 63.0%, while the corresponding Under-side rates were 69.0%, 70.3% and 85.5%.

The balanced side rate gives equal weight to the Over and Under conditional rates and is therefore less sensitive than the raw same-side rate to class imbalance. At 3.5 and 4.5, for example, the raw same-side rates are high because almost all observations fall on the xG-Under side, whereas the balanced side rates remain substantially lower. The balanced rate does not, however, solve sparse-sample instability: at the highest thresholds the xG-Over side contains very few observations. It is therefore retained as a descriptive class-balance summary rather than an inferential performance metric.

These percentages are **not betting returns** and are not measures of bookmaker mispricing. They contain no odds, margin, market timing or stake optimization. Their role is descriptive: they show how often cumulative xG and realized goals finish on the same side of an identical half-goal boundary while preserving the asymmetry between Over- and Under-side sample sizes.

5.6. Divergence by analytical band

Mean absolute band divergence increased across Bands 0–3:

$$0.41 \rightarrow 0.78 \rightarrow 0.85 \rightarrow 1.21.$$

The proportion with:

$$|\Delta_G| \geq 2$$

was:

$$6.9\%, \quad 11.6\%, \quad 20.7\%, \quad 29.4\%.$$

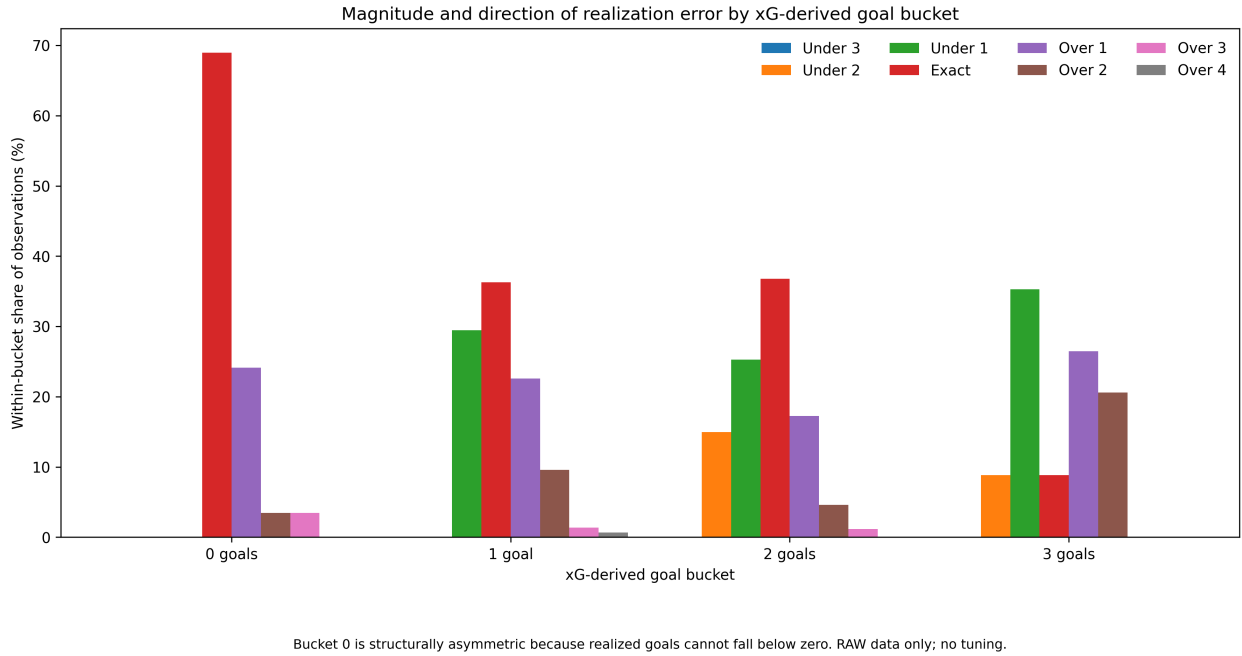


Figure 3: Figure 4. Magnitude and direction of realization divergence by half-goal-threshold xG analytical band.

Figure 4. Magnitude and direction of realization divergence by half-goal-threshold xG analytical band. Band 0 is structurally asymmetric because realized goals cannot fall below zero. The apparent increase in dispersion across higher bands is descriptive and requires replication on a larger sample.

5.7. League-specific distributions

Exact correspondence was observed in:

- Premier League: 28.8%;
- Serie A: 40.0%;
- La Liga: 40.0%;
- Bundesliga: 33.8%.

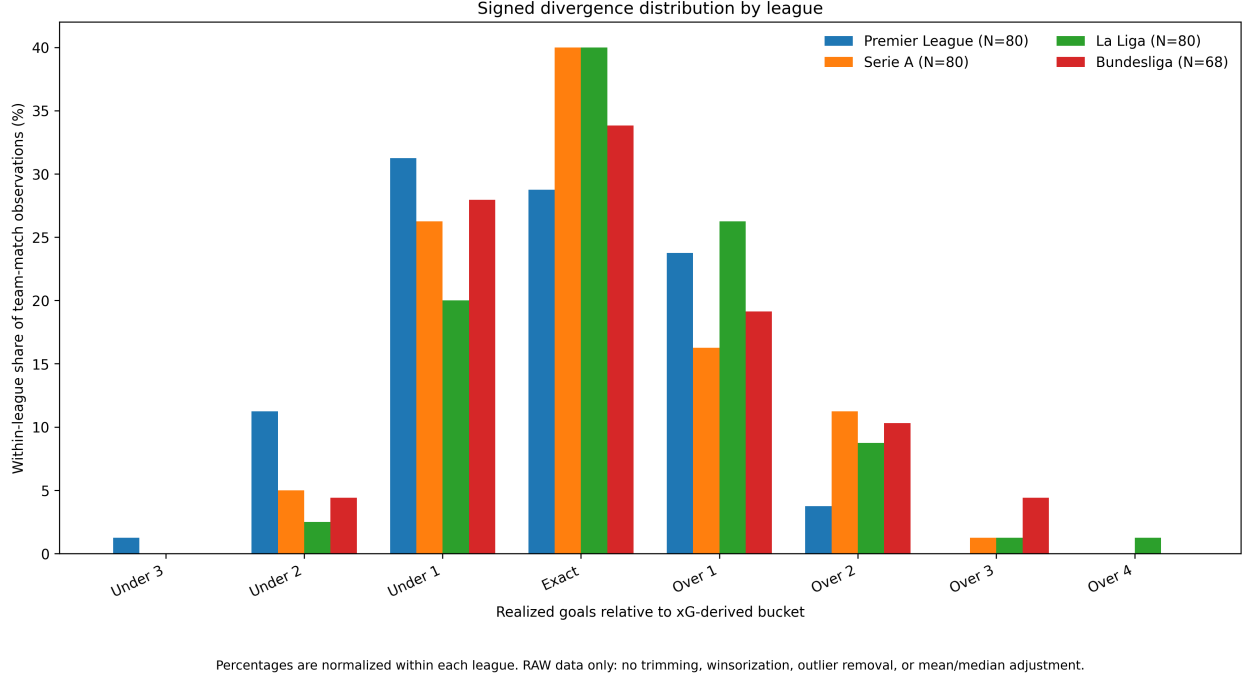


Figure 4: Figure 5. Realized-goal divergence from half-goal-threshold xG analytical bands by league.

The full signed distributions are shown in Figure 5.

Figure 5. Realized-goal divergence from half-goal-threshold xG analytical bands by league. Values are normalized within league. Differences are descriptive because league and xG provider are confounded in the present design.

The aggregate signed goal differences by league were:

$$-21, \quad +5, \quad +22, \quad +11$$

for Premier League, Serie A, La Liga and Bundesliga respectively.

Premier League and La Liga nearly cancel when pooled:

$$-21 + 22 = +1.$$

This provides a simple example of how aggregate balance can conceal different realization structures.

5.8. Empirical interpretation

The empirical results support three narrow conclusions.

First, mapping cumulative xG onto betting-motivated half-goal regions leaves substantial individual realization dispersion.

Second, that dispersion is not itself evidence of model failure. The didactic Poisson reference shows that a low exact-band frequency can arise naturally from a probabilistic count process.

Third, signed positive and negative deviations can offset one another:

$$\text{aggregate cancellation} \neq \text{individual realization certainty}.$$

The empirical contribution is therefore not an “accuracy score” for xG. It characterizes the relationship between a continuous model-implied expectation and discrete realized outcomes. The secondary line screen provides the corresponding binary half-goal view, while remaining deliberately separate from any claim about market value or profitability.

6. From Post-Match Comparison to Human Decision Environments

6.1. The post-match comparison concerns different statistical objects

A completed football score and cumulative xG are not competing measurements of the same quantity.

The final score is the observed realized outcome of the completed match. Cumulative xG is a conditional, model-dependent statistical representation of the registered shot opportunities generated from a finite information set.

Accordingly:

$$\text{realized outcome} \neq \text{conditional statistical representation}.$$

A large difference between the two does not by itself show that xG is invalid, nor does a small difference prove that the model was accurate at shot level.

The interpretive problem arises when the two objects are treated as interchangeable—for example, when cumulative xG is treated as the score that deterministically should have occurred, or when one divergent score is treated as sufficient evidence that the probabilistic representation had no analytical value.

6.2. Numerical precision as a cue

Research on advice taking shows that numerical precision can affect perceived advice quality and reliance even when underlying accuracy is controlled (Schultze & Loschelder, 2021).

Therefore a value such as 1.83 may be perceived as more authoritative than a coarser representation even though extra decimal places do not reveal aleatoric or epistemic uncertainty.

Whether this occurs specifically for xG remains an empirical question.

6.3. Information quantity and confidence

Hall et al. (2007) showed that additional sports-prediction information can increase confidence without an equivalent improvement in accuracy.

For the proposed experiment, the relevant question is therefore not merely whether statistical information changes confidence, but whether changes in confidence and wagering behaviour are accompanied by improved probabilistic judgment.

6.4. Source credibility and authority

The same statistical content can be interpreted differently depending on who produces and presents it.

A useful decomposition is:

$$\textit{Signal} = (\textit{Content}, \textit{Producer}, \textit{Carrier}, \textit{Presentation}).$$

For football statistics:

- **Content** = the numerical information, including xG where supplied;
- **Producer** = model/provider;
- **Carrier** = broadcaster, analytics platform or bookmaker;
- **Presentation** = match overlay or statistics-only interface.

Source-credibility research provides a basis for testing whether equivalent numerical information receives different weight under different presentation and source contexts (Sanz-Menéndez & Cruz-Castro, 2019; Hoogeveen et al., 2022).

6.5. Anchoring, availability and representativeness

Tversky and Kahneman’s (1974) framework identifies mechanisms that may become relevant in rapid numerical judgment.

A displayed statistic may act as an anchor. Recent high-danger actions may become especially available in memory. A sequence that visually resembles “domination” may be treated as representative of an imminent goal.

These are theoretical mechanisms, not established xG effects.

6.6. Perceived predictive control

Gambling research on perceived or illusory control provides a further testable mechanism (Dixon, 2000).

A bettor may interpret additional numerical information not merely as evidence but as an increase in the ability to predict an uncertain outcome.

The relevant discrepancy is between perceived predictive usefulness and demonstrated decision quality.

No claim is made that xG displays have already been shown to cause distorted betting behaviour. Existing evidence provides a rationale for testing how different information environments affect judgment and wagering.

The term *misleading* is therefore not treated here as an intrinsic property of xG. A valid analytical metric may become a misleading decision cue **in use** if its presentation causes users to infer greater certainty, completeness, event-specificity or prospective predictive value than the metric actually provides for the target decision. This is an empirical condition to test, not an assumption about bookmaker intent or about the inherent validity of xG.

6.7. Live xG is a backward-looking summary used in a prospective decision

At time t ,

$$XG_t = \sum_{j: time_j \leq t} \hat{p}_j$$

summarizes registered opportunities accumulated up to time t .

A live bet concerns a future outcome:

$$Y_{t+1:T}.$$

Therefore:

$$XG_t \neq P(Y_{t+1:T}).$$

Using cumulative xG in a live wagering decision requires an additional predictive inference from past match information to the future betting target. In particular, the cumulative value shown at time t is not itself:

$$P(\text{final Over 2.5} \mid \mathcal{J}_t),$$

$$P(\text{full-time 1X2} \mid \mathcal{J}_t),$$

or:

$$P(\text{next goal} \mid \mathcal{J}_t),$$

where $\mathcal{J}_t$ denotes the information available at the decision time.

This motivates a precise behavioural question:

When xG is displayed live, do users attribute more event-specific or prospective predictive authority to the number than is justified by its actual incremental predictive value for the betting target?

The present paper does not assume that such over-attribution occurs.

6.8. xG-specific weighting requires an xG-specific manipulation

The present stratified three-environment prospective design compares broader information environments. If the statistical displays contain several metrics, it cannot by itself identify how much decision weight is assigned specifically to xG.

An xG-specific claim about under-weighting or over-weighting would require a separate manipulation that varies xG presentation while holding the rest of the information package constant, together with an outcome-specific predictive benchmark.

Accordingly, the present experiment focuses on information-environment effects, probabilistic judgment and betting decisions rather than claiming to measure an isolated xG weight.

7. Proposed Prospective Betting Experiment

7.1. Objective and boundary with the retrospective analysis

The retrospective band and half-goal-line analyses motivate the problem of single-event interpretation, but they do **not** estimate the predictive value of xG for the three prospective betting outcomes.

The proposed live experiment is a distinct behavioural study. It compares decisions across three information environments and three participant strata at prespecified match checkpoints.

7.2. Participant strata

Participants are divided into three broad strata:

- **P1 — football experts:** adults with substantial football-domain expertise, with priority given to genuinely expert or professional profiles rather than student convenience samples;
- **P2 — regular bettors / football fans:** adults with recurring football and betting exposure;
- **P3 — football/betting novices:** adults with comparatively limited exposure to both football analytics and betting.

These strata are supplemented by continuous measures of:

- football expertise;
- statistical expertise;
- betting experience;
- xG familiarity.

Betting-history measures should include betting frequency, in-play frequency, cash-out use, loss-chasing behaviour and the Problem Gambling Severity Index (PGSI) or an equivalent validated measure (Miller et al., 2013).

Operational eligibility criteria and cutoffs for P1, P2 and P3 will be prespecified and preregistered before recruitment. The continuous expertise and experience measures will be retained even after categorical stratum assignment.

7.3. Stratified information-environment design

The participant strata are measured characteristics rather than randomized treatments. Within each stratum, participants are randomly assigned to one of three information environments, producing nine participant-by-environment cells. Randomization therefore supports causal interpretation of the information-environment contrast within strata; differences between experts, regular bettors and novices remain between-group comparisons rather than causal effects of expertise.

Test 1 — match visible, no analytical statistics

Participants watch the live match without analytical statistical overlays. Their judgment is based on direct visual observation of the football event and the ordinary match state.

Test 2 — match visible, statistics overlaid

Participants watch the same match while statistical information is shown in overlay. The statistical representation includes xG where it is supplied by the selected provider/interface.

Test 3 — statistics only

Participants do not watch the match video. Their judgments are based only on the statistical information supplied through the bookmaker/provider-style interface, including xG where available.

Test 2 and Test 3 use the same frozen statistical panel, with the same variables, provider source, update timing and numerical values at each checkpoint. The layouts should be made as presentation-equivalent as practicable; any unavoidable interface differences are treated as part of the broader information environment rather than as an isolated effect of video presence.

Within a given checkpoint and comparison, all non-environment information is held constant. Video and statistical feeds must also be synchronized to the same event state so that one environment does not receive information about a material match event earlier than another.

The design therefore compares **information environments**, not a within-person sequence in which xG is revealed after an initial judgment. There is one decision at each checkpoint under the participant's assigned information environment.

Because the statistical environments may contain multiple metrics, differences between Test 1 and Tests 2–3 should not automatically be interpreted as the isolated causal effect of xG alone. They quantify the effect of the broader statistical representation in which xG is a focal component.

The primary information-environment contrasts are prespecified as follows: Test 2 versus Test 1 estimates the effect of adding the statistical panel while direct match observation is retained; Test 3 versus Test 2 estimates the effect of removing direct match observation while holding the statistical panel constant; Test 3 versus Test 1 provides the broader contrast between direct match observation without analytical statistics and statistics-only decision-making.

7.4. Three distinct decision tasks at prespecified checkpoints

The experimental framework is applied to three distinct betting decisions:

- 15 minutes: final-match Over/Under 2.5;
- half-time: full-time 1X2;

- 60 minutes: next goal.

These observations are repeated within the broader experimental programme, but the target variable changes at each checkpoint. Differences between the 15-minute, half-time and 60-minute results therefore cannot be interpreted as a pure effect of match progression or elapsed time. Each checkpoint is analysed primarily as a distinct decision task, with cross-checkpoint comparisons treated as secondary.

7.5. Common decision structure

Before the test, each participant receives the same fixed experimental bankroll and is told that there are three predefined opportunities to use it: at 15 minutes, half-time and 60 minutes, each linked to the specified market for that checkpoint.

Allocation of the bankroll is left to the participant. The subject may use none, part or all of the available bankroll at any one checkpoint, may divide it across more than one checkpoint, and may choose not to bet at any stage. The timing and amount committed are therefore behavioural choices rather than fixed per-bet quantities.

When bankroll remains available, **NO BET** is a genuine voluntary option. If the bankroll has already been fully committed at an earlier checkpoint, the later checkpoint is coded separately as **no wager possible** — **bankroll exhausted**, rather than as a voluntary **NO BET**. Subjective probability, confidence and the non-wager perceptual measures are still collected at that checkpoint.

At each checkpoint the experiment records:

- subjective outcome probability;
- confidence;
- wagering state (**BET**, voluntary **NO BET**, or **BANKROLL EXHAUSTED**);
- selected market side where applicable;
- amount committed from the experimental bankroll;
- remaining bankroll as a state variable;
- response time;
- perceived credibility;
- perceived objectivity;
- perceived predictive usefulness;
- perceived predictive control.

Because bankroll allocation is endogenous to the participant’s decision, both bet incidence and amount committed can be analysed as behavioural outcomes. Remaining bankroll is treated as a state variable generated by previous allocation decisions, not as an independent behavioural response.

7.6. Minute 15 — full-match Over/Under 2.5

At the 15th minute, the target outcome is:

$$Y_{15} = \mathbf{1}(\text{final total goals} > 2.5).$$

Eligible scenarios must still have an unsettled Over/Under 2.5 outcome at the checkpoint.

Participants provide a subjective probability for Over 2.5 and then choose:

$$OVER, \quad UNDER, \quad NO \text{ BET},$$

with the participant choosing how much of the remaining experimental bankroll, if any, to commit.

7.7. Half-time — full-time 1X2

At half-time, the target is the final match result:

$$Y_{HT} \in \{Home, Draw, Away\}.$$

Participants provide a probability vector satisfying:

$$P(H) + P(D) + P(A) = 1$$

and then choose:

$$HOME, \quad DRAW, \quad AWAY, \quad NO \text{ BET}.$$

7.8. Minute 60 — next goal

At the 60th minute:

$$Y_{60} \in \{Home \text{ next}, Away \text{ next}, No \text{ further goal}\}.$$

Including **No further goal** avoids implicitly conditioning on another goal occurring.

Participants provide the complete probability vector, confidence and betting decision, including the amount of the remaining experimental bankroll, if any, to commit.

7.9. Outcome-specific predictive benchmark

The behavioural comparison should be interpreted alongside separate predictive benchmarks for the exact target at each checkpoint.

Minute 15

$$P(\text{final Over 2.5} \mid \text{information at 15'})$$

Half-time

$$P(H), P(D), P(A) \mid \text{information at HT}$$

Minute 60

$$P(\textit{Home next}), P(\textit{Away next}), P(\textit{No further goal}) \mid \textit{information at 60'}$$

To isolate the incremental predictive contribution of xG, the primary benchmark comparison must be **nested**: the two models compared for the xG contribution use the same historical and live information, differing only in whether xG is included.

A useful hierarchy is:

$$M_H = \text{historical} / \text{baseline information},$$

$$M_{H+L-xG} = \text{historical} + \text{live information excluding xG},$$

and:

$$M_{H+L+xG} = \text{historical} + \text{the same live information} + \text{xG}.$$

The primary xG increment is therefore the out-of-sample difference between M_{H+L-xG} and M_{H+L+xG} . Other model comparisons may be reported for context but do not replace this nested contrast.

All estimates of incremental predictive value must be obtained **out of sample** using a prespecified train/validation/test or cross-validation procedure.

Performance should be evaluated with proper scoring rules such as Brier score and log loss, with calibration assessed separately (Gneiting & Raftery, 2007).

7.10. Behavioural response and decision quality

The experiment distinguishes two outcome layers.

Behavioural response includes differences in:

- subjective probability;
- confidence;
- bet/no-bet frequency;
- selected side;
- amount committed from the experimental bankroll;
- response time.

State information includes remaining bankroll and whether a wager is still feasible at the check-point.

Decision quality includes differences in:

- Brier score;
- log loss;
- calibration;
- discrimination where appropriate.

The information environments may increase or decrease confidence, betting propensity or decision quality. These effects should be estimated directly rather than labelled in advance as under- or over-weighting of xG.

If multiple statistics are presented together, any observed environment effect must be attributed to the statistical information package unless an xG-specific manipulation is introduced in a later study.

7.11. Analysis plan

The repeated-checkpoint structure supports mixed-effects models with participant and match/scenario random effects.

Primary comparisons should estimate:

- information-environment effects within each participant stratum;
- participant-stratum differences within each information environment;
- information-environment $\times$ participant-stratum interactions;
- task-specific information-environment effects at 15', HT and 60'.

Candidate models include:

- mixed-effects regression for subjective probabilities and confidence;
- logistic or multinomial mixed-effects models for betting choice among checkpoints where bankroll remains available;
- models of bankroll allocation conditional on betting, with prior allocations and remaining bankroll represented explicitly;
- response-time models;
- calibration and proper-scoring comparisons between experimental cells.

Football expertise, statistical expertise, betting experience and xG familiarity should also be analysed as continuous moderators rather than inferred solely from P1/P2/P3 membership.

Primary outcomes, exclusions, checkpoint eligibility rules, sample-size determination and statistical models should be preregistered before data collection.

7.12. Ethics

Any implementation should include:

- adult informed consent;
 - appropriate safeguards for participants with gambling-related vulnerability;
 - no requirement to risk personal funds;
 - limited experimental stakes if incentives are used;
 - persistent NO BET;
 - debriefing;
 - prior ethical review.
-

8. Discussion

8.1. What the retrospective screening establishes

The empirical analysis does not establish that xG is inaccurate.

It establishes that after cumulative xG is projected onto half-goal-threshold analytical bands, individual realized goal counts remain distributed around those bands.

The observed exact-band rate of 35.7% should therefore not be interpreted as a failure rate.

The didactic Poisson comparison illustrates the point: a simple probabilistic count process with $\lambda = xG$ produces an average exact-band probability of approximately 34.2%. The comparison is not a calibration test and does not imply that the observed data are generated by a Poisson process.

The correct conclusion is:

A probabilistic expected value can be informative without identifying the realized goal count of a single team-match.

8.2. Realization variability is central to interpretation

A central source of uncertainty in this empirical section is realization variability. Even if the conditional expectation were known perfectly, individual outcomes would remain variable.

Post-match interpretation can fail in two directions without requiring a psychological claim about the user's weighting. A realized score can be treated as if it invalidated the statistical representation, or the xG value can be treated as if it were the deterministic score that should have occurred. Both readings collapse two different statistical objects into one.

8.3. Why half-goal boundaries matter

The half-goal discretization is not intended to recover the hidden output of an xG model.

It is designed to connect a continuous expected-value statistic with discrete boundaries that have operational meaning in football betting.

A 2.5-goal market has a binary settlement boundary:

$$0, 1, 2 \quad \text{versus} \quad 3, 4, 5, \dots$$

There is no realized state of 2.5 goals.

This illustrates the core transition examined by the paper: continuous probabilistic information must ultimately be converted into discrete actions and settlements when used for wagering.

The threshold-sensitivity analysis indicates that the principal descriptive result is not an artefact of a small number of observations immediately adjacent to the chosen cutpoints.

The secondary two-sided threshold screen provides the more direct betting-boundary analogue. At 0.5, 1.5 and 2.5, the Over-side realization rates were 79.9%, 71.4% and 63.0%, while the corresponding Under-side rates were 69.0%, 70.3% and 85.5%. These are threshold-side realization frequencies only, not evidence of betting profitability.

8.4. Aggregate cancellation and single-event decisions

The signed mean divergence is only:

$$+0.055.$$

Yet exact band correspondence is 35.7%.

This is not contradictory. Positive and negative realizations cancel in aggregate.

The point matters because a betting decision is not made on the league-average residual. It is made on one match and one market.

Therefore:

$$\text{aggregate balance} \neq \text{single-event certainty}.$$

8.5. Model dependence remains distinct from aleatoric uncertainty

The paper separates two issues that are often conflated.

First:

$$\text{aleatoric outcome variation}.$$

Second:

$$\text{epistemic/model dependence}.$$

Provider disagreement, annotation choices and information-set differences concern the second. Realization dispersion around a fixed expectation concerns the first.

The present empirical screening characterizes realized dispersion that is **consistent with aleatoric variation**, but it does not decompose observed divergence into aleatoric and epistemic components. A systematic same-match, multi-provider study would address model/provider dependence and help quantify one important source of epistemic variation.

8.6. Post-match analysis and live betting are different uses of the same statistic

Post-match xG is retrospective. In live use, cumulative xG is better described as a backward-looking summary of registered opportunities accumulated up to the current match state.

Live betting is prospective. The user must forecast a future target such as:

$$\text{final Over 2.5,}$$

$$\text{full-time 1X2,}$$

or:

next goal.

Cumulative live xG can be informative for these targets, but it is not itself the required probability. An additional predictive mapping is necessary.

This is where the information environment and the quality of human probabilistic judgment become the central behavioural questions.

8.7. What existing in-play betting evidence does and does not establish

Existing in-play betting research provides evidence that wagering behaviour responds to evolving match information and that bettors use live match cues (Killick & Griffiths, 2021; Michels et al., 2023; Ötting et al., 2024). Other gambling research documents mechanisms such as tilting, inducement effects and subjective-probability distortions in sports-betting contexts (Torrance et al., 2022; Ó Ceallaigh et al., 2025; Chegere et al., 2025).

This literature does **not** establish that displaying xG specifically causes distorted betting behaviour. The proposed experiment tests the broader causal effect of information environment within participant strata; it does not isolate an xG-specific causal effect when multiple statistics are displayed together.

8.8. Decision effects should be interpreted alongside incremental predictive value

The relevant question is not simply whether bettors react to statistical information. The experiment asks whether judgment quality and wagering behaviour differ across match-only, match-plus-statistics and statistics-only environments, and whether those differences depend on expertise.

A separate predictive benchmark then determines how much incremental predictive information xG contributes to the exact target outcome. Because the statistical panels may contain multiple metrics, the behavioural environment effect and the xG-specific model contribution must remain analytically distinct.

8.9. Evidence status

Component	Status in this paper	Supported conclusion	Not established
Conditional xG framework	Methodological/statistical	xG depends on information set, model and data	Complete event-specific probability
Team-match band screening	Original descriptive evidence	Realization dispersion around half-goal-derived bands	Shot-level calibration failure
Poisson reference	Didactic probabilistic comparison	Low exact-band rates can arise from count variability	True football goal-generating process

Component	Status in this paper	Supported conclusion	Not established
Threshold sensitivity	Original robustness check	Main descriptive pattern is not boundary-adjacent only	Invariance to every possible transformation
Two-sided half-goal-threshold screen	Original secondary descriptive evidence	Over- and Under-side realization frequencies differ across thresholds	Betting profitability, bookmaker mispricing or calibration
Broader physical event state	Source-supported conceptual argument	Richer model state can reduce representational limits, but player, goalkeeper, ball, surface and environmental state remain model-specific conditioning questions	Cause of observed Δ_G , physical-variable effect sizes, or universal omission by modern xG models
Information-environment mechanisms	Theory-supported hypotheses	Plausible mechanisms for testing	xG-specific causal effect
Proposed betting experiment	Stratified three-environment prospective design	Testable randomized information-environment effects within participant strata and descriptive/moderated differences across strata	No behavioural result yet; no isolated xG causal effect if multiple statistics are shown together

8.10. Limitations

Several limitations constrain interpretation.

First, the empirical sample contains only the first four rounds of four leagues.

Second, provider and league are confounded.

Third, the analysis uses cumulative team-match xG rather than shot-level probabilities and cannot formally test provider calibration.

Fourth, the 308 team-match observations arise from 154 fixtures and are therefore paired within matches rather than fully independent. The present analysis is descriptive; any future inferential analysis should account for match-level clustering.

Fifth, the betting-motivated half-goal transformation is an analytical construction. It is motivated by betting-market settlement boundaries but is not a native xG representation.

Sixth, the Poisson reference is deliberately simplified. A shot-level Poisson-binomial or dependence-aware count model would provide a more structurally faithful probabilistic benchmark if all constituent shot probabilities were available.

Seventh, the current study does not estimate the incremental predictive value of xG for Over/Under 2.5, 1X2 or next-goal markets. Those are targets of the proposed prospective programme.

Eighth, no behavioural experiment has yet been conducted. No causal conclusion about the effect of the information environments on betting behaviour is claimed in the present study.

Ninth, if the statistical display contains multiple metrics, differences between the experimental information environments cannot be attributed exclusively to xG without an additional xG-specific manipulation.

Tenth, some source values differ in reported decimal precision. Precision flags and provenance should remain in the supplementary material.

8.11. Research programme

The next empirical stages are:

1. expand the retrospective dataset across more rounds and seasons;
2. reconstruct shot-level data where possible to replace the didactic Poisson reference with a more appropriate count distribution;
3. perform a prespecified same-match cross-provider robustness study;
4. reconstruct or capture live match states at 15', HT and 60' under the three frozen information environments;
5. estimate the out-of-sample incremental predictive value of xG for the three specified betting outcomes;
6. preregister and conduct the stratified three-environment experiment across the three participant strata.

9. Conclusion

Expected Goals are probabilistic estimates. Their usefulness does not depend on the realized score matching the nearest integer implied by the cumulative value.

The present screening examined 308 team-match observations after mapping continuous cumulative xG onto goal regions separated by half-goal analytical boundaries. Exact band correspondence occurred in 35.7% of observations, while 16.6% lay at least two goals away. Mean absolute band divergence was 0.834 goals, yet signed mean divergence was only +0.055.

A didactic Poisson reference with $\lambda = xG$ produced an expected exact-band rate of approximately 34.2%. This comparison reinforces the principal statistical point:

$$\text{realization divergence} \neq \text{model failure.}$$

Individual outcomes remain subject to aleatoric variation even when a probabilistic expectation is informative.

In post-match analysis, the central interpretive error is to treat the realized score and cumulative xG as if they were interchangeable measurements of the same quantity. The score is the realized outcome; xG is a conditional statistical representation of the opportunities encoded by the model.

In live betting, the issue becomes prospective. Cumulative xG describes past registered opportunities, while the wager concerns a future outcome. The proposed experiment therefore randomizes match-only, match-plus-statistics and statistics-only information environments within expert, regular-bettor and novice strata, while a separate out-of-sample model benchmark estimates the incremental predictive contribution of xG. The three checkpoints use different target outcomes and are analysed primarily as distinct decision tasks rather than as a pure time-progression sequence.

The proposed experiment operationalizes this question at three checkpoints:

$15' \rightarrow \text{final Over/Under } 2.5,$

$HT \rightarrow \text{full-time } 1X2,$

$60' \rightarrow \text{next goal}.$

The broader conclusion is therefore:

xG is a conditional probabilistic representation rather than an observed match outcome. Its relationship with a single match result is subject to realization variability, and its use in live betting should be evaluated within the information environment in which human decisions are actually made.

This claim does not reject xG. Increasing the informational richness of contemporary xG models is a genuine methodological advance and can reduce representational limitations. It does not, however, remove the distinction between a model-conditioned probability estimate, the realized outcome of a single event, and the prospective probability of a future betting target. The paper therefore specifies the boundary between statistical expectation, realized outcome and human interpretation.

Data Availability

The retrospective screening was conducted from league-specific xG data sources recorded in the project workbooks. The v1.0 analysis includes threshold-sensitivity, didactic Poisson-reference and two-sided half-goal-threshold screening tables derived from the frozen xG-complete sample.

Provider-originating raw values will be redistributed only where the relevant source terms permit redistribution. If raw redistribution is restricted, the publication package will instead contain derived analytical tables, transformation rules, provenance metadata, source-gap and precision audits, reproducibility documentation and checksums sufficient to reconstruct the analysis from lawfully obtained source data.

No supplementary dataset should be considered final until the manuscript version, redistribution-rights review and associated checksums are frozen at publication.

Ethics Statement

The present study contains no human-participant experiment. The behavioural protocol described in Section 7 is a proposed future study only. Any implementation involving participants, monetary incentives or gambling-related decisions would require prior ethical review and appropriate safeguards.

Conflict of Interest

The author declares no conflict of interest relevant to the present analysis.

Funding

No external funding was received for the present work.

Supplementary Materials Planned for Publication

The final publication package is intended to include:

- league-level raw xG screening workbooks only where redistribution rights permit;
- consolidated 154-match / 308-observation analytical workbook;
- Δ_G distribution analysis;
- half-goal-threshold band 0–3 analysis;
- figure-source tables;
- threshold-sensitivity table;
- didactic Poisson-reference table;
- pooled and league-level two-sided half-goal-threshold screening table with balanced side rates;
- source and precision audit;
- methodology/readme document;
- frozen checksums for all deposited supplementary files.

These materials are not yet treated as the final archived supplementary package.

References

- Anzer, G., & Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*, 3, 624475. <https://doi.org/10.3389/fspor.2021.624475>
- Bandara, I., Shelyag, S., Rajasegarar, S., Dwyer, D., Kim, E.-j., & Angelova, M. (2024). Predicting goal probabilities with improved xG models using event sequences in association football. *PLOS ONE*, 19(10), e0312278. <https://doi.org/10.1371/journal.pone.0312278>

- Betfair. (2026). *Mercati calcio: regole di refertazione Under/Over e mercati goal*. Betfair Italia. <https://www.betfair.it/aboutUs/mercati/> (accessed 23 September 2026).
- Brechot, M., & Flepp, R. (2020). Dealing with randomness in match outcomes: How to rethink performance evaluation in European club football using expected goals. *Journal of Sports Economics*, 21(4), 335–362. <https://doi.org/10.1177/1527002519897962>
- Cefis, M., & Carpita, M. (2025). Accuracy and explainability of statistical and machine learning xG models in football. *Statistics*, 59(2), 426–445. <https://doi.org/10.1080/02331888.2024.2445305>
- Chegere, M. J., Falco, P., Nieddu, M. G., Pandolfi, L., & Stein, M. R. (2025). The magic of the game: Experimental evidence on sports betting behavior. *Journal of Economic Behavior & Organization*, 240, 107301. <https://doi.org/10.1016/j.jebo.2025.107301>
- Dambroz, F., Clemente, F. M., & Teoldo, I. (2022). The effect of physical fatigue on the performance of soccer players: A systematic review. *PLOS ONE*, 17, e0270099. <https://doi.org/10.1371/journal.pone.0270099>
- Dixon, M. R. (2000). Manipulating the illusion of control: Variations in gambling as a function of perceived control over chance outcomes. *The Psychological Record*, 50(4), 705–719. <https://doi.org/10.1007/BF03395379>
- Fédération Internationale de Football Association (FIFA). (2021). *FIFA Quality Programme for Natural Playing Surfaces*. <https://inside.fifa.com/innovation/standards/natural-playing-surfaces/fifa-quality-programme-for-natural-playing-surfaces> (accessed 24 September 2026).
- Fédération Internationale de Football Association (FIFA). (n.d.). *FIFA Quality Programme for Football Turf*. <https://inside.fifa.com/innovation/standards/football-turf/fifa-quality-programme-for-football-turf> (accessed 24 September 2026).
- Fernández, J., Bornn, L., & Cervone, D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, 110, 1389–1427. <https://doi.org/10.1007/s10994-021-05989-6>
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>
- Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes*, 103, 277–290. <https://doi.org/10.1016/j.obhdp.2007.01.003>
- Hong, S., Goff, J. E., & Asai, T. (2019). Effect of a soccer ball’s surface texture on its aerodynamics and trajectory. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, 233(1), 67–74. <https://doi.org/10.1177/1754337118794561>
- Hoogeveen, S., Haaf, J. M., Bulbulia, J. A., Ross, R. M., McKay, R., Altay, S., Bendixen, T., Berniūnas, R., Cheshin, A., Gentili, C., Georgescu, R., Gervais, W. M., Hagel, K., Kavanagh, C., Levy, N., Neely, A., Qiu, L., Rabelo, A., Ramsay, J. E., ... van Elk, M. (2022). The Einstein effect provides global evidence for scientific source credibility effects and the influence of religiosity. *Nature Human Behaviour*, 6, 523–535. <https://doi.org/10.1038/s41562-021-01273-8>
- Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110, 457–506. <https://doi.org/10.1007/s10994-021-05946-3>

- Ibrahim, R., Kingma, I., de Boode, V., Faber, G. S., & van Dieën, J. H. (2020). Angular velocity, moment, and power analysis of the ankle, knee, and hip joints in the goalkeeper's diving save in football. *Frontiers in Sports and Active Living*, 2, 13. <https://doi.org/10.3389/fspor.2020.00013>
- Illmer, S., & Daumann, F. (2022). The effects of weather factors and altitude on physical and technical performance in professional soccer: A systematic review. *JSAMS Plus*, 1, 100002. <https://doi.org/10.1016/j.jsampl.2022.100002>
- International Football Association Board (IFAB). (n.d.). *Law 2 - The Ball*. <https://www.theifab.com/laws/latest/tball/> (accessed 24 September 2026).
- Killick, E. A., & Griffiths, M. D. (2021). Why do individuals engage in in-play sports betting? A qualitative interview study. *Journal of Gambling Studies*, 37(1), 221–240. <https://doi.org/10.1007/s10899-020-09968-9>
- Lago-Peñas, C., Feng, R., Zhang, Y., Dong, W., Gong, B., & Gómez-Ruano, M.-Á. (2026). A probabilistic and dynamic reformulation of expected goals (xG): Methodological advances and modelling perspectives. *Intelligent Sports and Health*, 2(3), 151–154. <https://doi.org/10.1016/j.ish.2026.05.001>
- Lees, A., Asai, T., Andersen, T. B., Nunome, H., & Sterzing, T. (2010). The biomechanics of kicking in soccer: A review. *Journal of Sports Sciences*, 28, 805–817. <https://doi.org/10.1080/02640414.2010.481305>
- Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE*, 18, e0282295. <https://doi.org/10.1371/journal.pone.0282295>
- Michels, R., Ötting, M., & Langrock, R. (2023). Bettors' reaction to match dynamics: Evidence from in-game betting. *European Journal of Operational Research*, 310(3), 1118–1127. <https://doi.org/10.1016/j.ejor.2023.04.006>
- Miller, N. V., Currie, S. R., Hodgins, D. C., & Casey, D. (2013). Validation of the Problem Gambling Severity Index using confirmatory factor analysis and Rasch modelling. *International Journal of Methods in Psychiatric Research*, 22(3), 245–255. <https://doi.org/10.1002/mpr.1392>
- Ó Ceallaigh, D., Timmons, S., Robertson, D. A., & Lunn, P. D. (2025). Effects of inducements on sports gambling and decision-errors: An experimental study. *Journal of Behavioral Addictions*, 14(2), 959–971. <https://doi.org/10.1556/2006.2025.00056>
- Opta Analyst. (n.d.). *What is Expected Goals (xG)?* <https://theanalyst.com/articles/what-is-expected-goals-xg> (accessed 24 September 2026).
- Ötting, M., Michels, R., Langrock, R., & Deutscher, C. (2024). Demand for live betting: An analysis using state-space models. *Applied Stochastic Models in Business and Industry*, 40(2), 527–541. <https://doi.org/10.1002/asmb.2836>
- Peng, W., Zhang, Y., Wang, D., Sarmento, H., Vilas-Boas, J. P., & Ribeiro, J. (2026). Exercise-induced fatigue and soccer kicking performance: A systematic review and meta-analysis of performance outcomes and contextual moderators. *Frontiers in Physiology*, 17, 1818415. <https://doi.org/10.3389/fphys.2026.1818415>
- Pereira, E., Peek, K., McLean, C., Lavender, A. P., Ma'ay, F., Davey, P., Morris, S., & Georgieva, J. (2024). The effect of ball pressure and maximal isometric neck strength on head acceleration during purposeful heading in adult football players during heading drills. *JSAMS Plus*, 3, 100051. <https://doi.org/10.1016/j.jsampl.2023.100051>

- Sanz-Menéndez, L., & Cruz-Castro, L. (2019). The credibility of scientific communication sources regarding climate change: A population-based survey experiment. *Public Understanding of Science*, 28(5), 534–553. <https://doi.org/10.1177/0963662519840946>
- Scholtes, A., & Karakuş, O. (2024). Bayes-xG: Player and position correction on expected goals (xG) using Bayesian hierarchical approach. *Frontiers in Sports and Active Living*, 6, 1348983. <https://doi.org/10.3389/fspor.2024.1348983>
- Schultze, T., & Loschelder, D. D. (2021). How numeric advice precision affects advice taking. *Journal of Behavioral Decision Making*, 34(3), 303–310. <https://doi.org/10.1002/bdm.2211>
- Silva, D. C. F., Santos, R., Vilas-Boas, J. P., Macedo, R., Montes, A. M., & Sousa, A. S. P. (2017). Influence of cleats-surface interaction on the performance and risk of injury in soccer: A systematic review. *Applied Bionics and Biomechanics*, 2017, 1305479. <https://doi.org/10.1155/2017/1305479>
- Smith, M. R., Zeuwts, L., Lenoir, M., Hens, N., De Jong, L. M. S., & Coutts, A. J. (2016). Mental fatigue impairs soccer-specific decision-making skill. *Journal of Sports Sciences*, 34(14), 1297–1304. <https://doi.org/10.1080/02640414.2016.1156241>
- Torrance, J., Roderique-Davies, G., Greville, J., O’Hanrahan, M., Davies, N., Sabolova, K., & John, B. (2022). Conceptualising emotional and cognitive dysregulation amongst sports bettors: An exploratory study of ‘tilting’ in a new context. *PLOS ONE*, 17(2), e0264000. <https://doi.org/10.1371/journal.pone.0264000>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- xGStat. (2026). *Expected Goals (xG): Definition, Values and Limits*. <https://docs.xgstat.com/learn/expected-goals> (reviewed 18 September 2026; accessed 24 September 2026).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SUPPLEMENTAGGREGATEV1.xlsx](#)