
# Efficient and Reliable Estimation of Knowledge Graph Accuracy

Stefano Marchesin
University of Padua, Padua, Italy
stefano.marchesin@unipd.it

Gianmaria Silvello
University of Padua, Padua, Italy
gianmaria.silvello@unipd.it

## ABSTRACT

Data accuracy is a central dimension of data quality, especially when dealing with Knowledge Graphs (KGs). Auditing the accuracy of KGs is essential to make informed decisions in entity-oriented services or applications. However, manually evaluating the accuracy of large-scale KGs is prohibitively expensive, and research is focused on developing efficient sampling techniques for estimating KG accuracy. This work addresses the limitations of current KG accuracy estimation methods, which rely on the Wald method to build confidence intervals, addressing reliability issues such as zero-width and overshooting intervals. Our solution, rooted in the Wilson method and tailored for complex sampling designs, overcomes these limitations and ensures applicability across various evaluation scenarios. We show that the presented methods increase the reliability of accuracy estimates by up to two times when compared to the state-of-the-art while preserving or enhancing efficiency. Additionally, this consistency holds regardless of the KG size or topology.

**PVLDB Reference Format:**
Stefano Marchesin and Gianmaria Silvello. Efficient and Reliable Estimation of Knowledge Graph Accuracy. PVLDB, 17(9): 2392 - 2404, 2024.
doi:10.14778/3665844.3665865

**PVLDB Artifact Availability:**
The source code, data, and/or other artifacts are available at https://github.com/KGAccuracyEval/reliable-kg-estimation.

## 1 INTRODUCTION

In recent years, there has been a notable upsurge in the development of extensive Knowledge Graphs (KGs) encompassing millions of relational facts, primarily represented as triples in the form of subject-predicate-object $(s, p, o)$ relationships. Prominent examples of such KGs are Wikidata [37], DBpedia [2], YAGO [16], NELL [27], and DisGeNET [33]. However, current processes for constructing KGs are not flawless, resulting in a high degree of sparsity within the graph and the incorporation of numerous inaccuracies and wrong facts [10, 31]. As a result, evaluating the KG accuracy holds pivotal importance, as it serves multiple essential purposes, such as triggering the refinement of the construction process, gaining insights into the quality of the data, and providing valuable information for downstream applications.

KG accuracy evaluation has broad implications for databases [18, 19, 28], but also for search, recommender, and question answering systems [34, 35]. As also highlighted by industrial applications like Saga [19], the information in the KG needs to be correct to ensure an engaging user experience for entity-oriented services – making the on-demand evaluation of the KG accuracy a critical feature for any knowledge platform.

Evaluating KG accuracy requires annotating facts with correctness labels. However, obtaining high-quality labels to perform KG evaluation is costly, involving manual annotation by experts or crowdsourcing workers [13, 30]. Besides, annotating every fact in a large-scale KG is infeasible [32]. Recently, some studies have considered these challenges [13, 30, 32], framing the evaluation of KG accuracy as a constrained minimization problem [13, 32]. To tackle it, they employ iterative procedures that encompass sampling strategies to facilitate efficient data collection, estimators for ascertaining accuracy, and Confidence Intervals (CIs) to evaluate the robustness of these estimations. These procedures rely on the Wald method for constructing CIs [5], a widely recognized and used technique based on normal approximation [4]. However, the assessment of KG accuracy requires the estimation of *binomial proportions*, which denotes the ratio between the number of correct facts (successes) and the total number of facts. In this context, Wald intervals are limited by zero-width and overshooting intervals, affecting the reliability of the estimations [38] (see Figure 2). These issues are particularly pronounced when the proportion approaches the boundaries [4], that is, values close to zero or one, which are fairly common in real-world scenarios [10, 31].

Thus, when designing approaches for KG evaluation, we need efficient methods that ensure low annotation costs while providing reliable CIs that account for the binomial properties of KG accuracy.

In this regard, our **contributions** are as follows. First, we highlight the problems of current state-of-the-art solutions for KG accuracy estimation based on the Wald interval by validating its limitations on real-life and synthetic KGs.

Secondly, we introduce and evaluate a family of binomial CIs that overcome Wald limitations, providing reliable CIs for KG accuracy estimation. We conduct theoretical and empirical comparisons between CIs, identifying the Wilson interval [42] as the method providing the best trade-off between efficiency and reliability.

Building on this result, we provide solutions to adjust Wilson and the other binomial intervals to complex but widely used sampling designs, such as clustering and stratification. Experiments on various KGs with different accuracy and topology underscore the importance of these adjustments to avoid incurring unstable CIs.

To further validate the robustness of the solutions based on Wilson intervals, we compare them to the state-of-the-art solutions by Gao et al. [13], which adopt Two-stage Weighted Cluster Sampling (TWCS) alongside Wald intervals. The results show that Wilson is up to two times more reliable when Wald falls short and more efficient when both intervals are reliable, confirming that Wilson should be used instead of Wald for KG accuracy evaluation.


Proceedings of the VLDB Endowment, Vol. 17, No. 9 ISSN 2150-8097.
doi:10.14778/3665844.3665865

Finally, to assess the scalability of the proposed solutions, we experiment on a synthetic KG comprising more than 100 million triples. The findings indicate that performance is independent of KG size and topology, maintaining high reliability at low costs.

***Outline.*** Section 2 introduces the problem and evaluation framework. Section 3 outlines state-of-the-art sampling designs and estimators. Section 4 outlines Wald, the binomial intervals, and the theoretical advantages of Wilson. Section 5 shows how to adjust these intervals to complex sampling designs. Sections 6 and 7 present experimental setup and results. Section 8 reviews related work. Section 9 concludes the paper.

## 2 PROBLEM AND FRAMEWORK

In this section, we first present the essential background and the required notation. Then, we introduce the problem and its formulation. Finally, we outline the evaluation framework.

### 2.1 Preliminaries

A KG is a directed, edge-labeled multi-graph [3], usually defined as $G = (V, R, \eta)$, where $V = \{E \cup A \cup B\}$ is the set of nodes in $G$, where $E$ are entities, $A$ attributes, and $B$ blank nodes; $R$ is the set of relationships between nodes in $G$; and $\eta : R \rightarrow (E \cup B) \times (E \cup A \cup B)$ is a function assigning an ordered pair of nodes to each relationship. The $\eta$ function produces the ternary relation $T$ of $G$ [3]. This work considers ground RDF graphs (i.e., without blank nodes), hence $\eta : R \rightarrow E \times (E \cup A)$. Thus, the ternary relation $T$ is the set of $(s, p, o)$ triples such that $s \in E$, $p \in R$, and $o \in E \cup A$, where $M = |T|$ is its size. Triples whose object is an entity are called triples with entity property, whereas those with attribute objects are known as triples with data property. A triple is also a fact; the two terms are used interchangeably. This work considers triples as first-class citizens next to nodes and relationships. Therefore, we define a KG as $G = (V, R, T, \eta)$, and an *entity cluster* $G[e] = \{(s, p, o) \in T \mid s = e\}$ as a set of triples in $T \in G$ sharing the same subject $e \in V$.

### 2.2 Problem Formulation

Let us denote the correctness of a triple $t \in T$ by an indicator function $\mathbb{1}_T(t) \rightarrow \{0, 1\}$, where 1 indicates correctness and 0 incorrectness. Then, the KG accuracy can be defined as the mean accuracy of its triples:

$$\mu(G) = \frac{\sum_{t \in T} \mathbb{1}_T(t)}{M} \tag{1}$$

where the value of $\mathbb{1}_T(t)$ is computed by manual annotation. Note that we consider correctness a binary problem as in triple validation [11], given that an atomic fact is either correct or incorrect.

Given that manually evaluating every triple of a large-scale KG to assess its accuracy is infeasible, the common practice is to estimate $\mu(G)$ with an estimator $\hat{\mu}$ calculated over a relatively small sample drawn according to a sampling strategy $\mathcal{S}$ designed to select $T_\mathcal{S} \subset T$ triples to annotate. The result is a sample $G_\mathcal{S} = (V_\mathcal{S}, R_\mathcal{S}, T_\mathcal{S}, \eta)$, where $V_\mathcal{S} \subset V$ and $R_\mathcal{S} \subset R$. To evaluate the accuracy of $G$, the estimator $\hat{\mu}$ must be unbiased; that is, $E[\hat{\mu}] = \mu(G)$. Moreover, $\hat{\mu}$ is a point estimator; hence, a $1 - \alpha$ CI at a given significance level $\alpha$ has to be provided to quantify the uncertainties in the sampling procedure. The larger the size of the sample $G_\mathcal{S}$, the smaller the width of the CI, until obtaining a CI of zero width when the sample $G_\mathcal{S}$ is equivalent to $G$ itself. A relevant measure associated with CIs is the Margin of Error (MoE), which is the half-width of a CI.

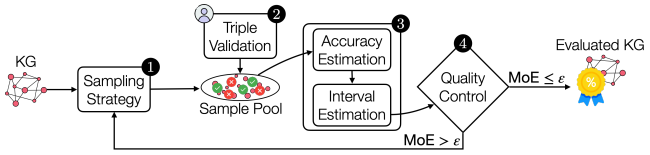


**Figure 1: Efficient KG Accuracy Evaluation Framework.**

Now, let $G_\mathcal{S} = \mathcal{S}(G)$ be a sample drawn using a sampling design $\mathcal{S}$, and $\hat{\mu}$ be an estimator of $\mu(G)$ based on $G_\mathcal{S}$. Let $\text{cost}(G_\mathcal{S})$ be a function denoting the cost of manually evaluating the correctness of the elements in $G_\mathcal{S}$. Following Gao et al. [13], we can define KG accuracy evaluation as a constrained minimization problem.

**Problem Formulation.** Given a KG $G$ and an upper bound $\varepsilon$ for the MoE of a $1 - \alpha$ CI:

$$\begin{aligned} \underset{\mathcal{S}}{\text{minimize}} \quad & E[\text{cost}(\mathcal{S}(G))] \\ \text{subject to} \quad & E[\hat{\mu}] = \mu(G), \text{MoE}(\hat{\mu}, \alpha) \leq \varepsilon \end{aligned}$$

To be optimized, the problem requires a sampling strategy $\mathcal{S}$ that minimizes the cost associated with manually evaluating elements in the sample. At the same time, the problem imposes a constraint on the MoE, which must be kept below an upper bound $\varepsilon$. The problem remains unsolved until the CI is sufficiently short. Hence, the CI is crucial in optimizing the problem, working in tandem with the sampling strategy. CIs with faster convergence enable the problem to be satisfied with smaller samples, thereby reducing the number of required annotations. However, it is crucial for CIs also to be reliable, ensuring they encompass the true KG accuracy $1 - \alpha$ times, or approximately so [1]. This entails a trade-off between desired efficiency and required reliability in CI selection.

Despite its significance, the impact of CIs has been overlooked in prior research, which predominantly focused on choosing an optimal sampling strategy [13, 32]. This research addresses this crucial gap.

### 2.3 Evaluation Framework

The minimization problem is optimized via an evaluation framework that works as an iterative procedure, whose four phases are depicted in Figure 1.

In ❶, a small batch of samples is collected from the KG under a given sampling strategy $\mathcal{S}$. The sampling strategy should be chosen to minimize the cost function, which is the optimization objective (see Section 3). In ❷, (manual) annotations are acquired for the new samples and stored together with previous ones. In ❸, given the accumulated annotations and the sampling design, an estimator computes an unbiased estimate $\hat{\mu}$ of the KG accuracy. The estimate is then used to build the corresponding $1 - \alpha$ CI, providing fast convergence and high reliability (see Section 4). In ❹, a quality control stage monitors if the generated CI satisfies the given upper bound, meaning if $\text{MoE}(\hat{\mu}, \alpha) \leq \varepsilon$. If the criterion is met, the process is halted, and the accuracy estimate and corresponding CI are reported. Otherwise, the process loops back to ❶.

This framework samples and estimates iteratively and stops as soon as the CI satisfies the specified threshold $\varepsilon$. This way, the

**Table 1: Notation.**

| | |
|---|---|
| $N$ | Number of entities in $G$ |
| $M_i$ | Size of the $i$th entity cluster |
| $M = \sum_{i=1}^{N} M_i$ | Total number of triples in $G$ |
| $n$ | Number of entity clusters in the sample |
| $m$ | Number of triples drawn in each cluster |
| $\tau_i$ | Number of correct triples in the $i$th cluster |
| $\mu_i = \tau_i / M_i$ | Accuracy of the $i$th cluster |

approach prevents oversampling and unnecessary manual annotations, providing accurate estimations while minimizing costs.

## 3 SAMPLING DESIGN: STATE OF THE ART

The strategies that have been proposed to perform efficient KG accuracy evaluation [13, 32] adopt established sampling techniques and estimators [7]. We present them along with their unbiased estimators. Table 1 reports frequently used notation.

### 3.1 Simple Random Sampling

Simple Random Sampling (SRS) draws a sample of $n_T$ triples from $G$ without replacement. If the KG is large, then the probability of choosing one triple twice is low, and we can use sampling with replacement instead [5]. Under SRS, the unbiased estimator of $\mu(G)$ is the sample proportion

$$\hat{\mu}_s = \frac{1}{n_T} \sum_{i=1}^{n_T} \mathbb{1}_T(t_i) \tag{2}$$

and the corresponding estimation variance is $V(\hat{\mu}_s) = \frac{\hat{\mu}_s(1-\hat{\mu}_s)}{n_T}$.

### 3.2 Cluster Sampling

Cluster sampling offers an efficient alternative for data sampling when dealing with large KGs [13, 32]. We first introduce Weighted Cluster Sampling (WCS) and then present TWCS and its estimator.

WCS draws $n$ entity clusters with probabilities $\pi_i$ proportional to their sizes. Denoting the cardinality of the $i$th cluster as $M_i = |G[e_i]|$, the cluster probability can be written as $\pi_i = M_i/M$. WCS is a single-stage sampling design, meaning all triples in the sampled clusters are manually evaluated. When clusters are large, as in most KGs, the cost of WCS may be prohibitive.

To overcome this limitation, TWCS can be used. TWCS consists of two stages: (1) sample entity clusters with WCS; (2) from each $i$th sampled cluster, sample $\min\{M_i, m\}$ triples with SRS without replacement. By denoting the (estimated) mean accuracy of the sampled triples in the $i$th cluster as $\hat{\mu}_i$, the unbiased estimator of $\mu(G)$ under TWCS is

$$\hat{\mu}_{w,m} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_i \tag{3}$$

with estimation variance $V(\hat{\mu}_{w,m}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\mu}_i - \hat{\mu}_{w,m})^2$.

### 3.3 Stratified Sampling

Stratified Sampling (SS) stratifies entity clusters into $Q$ non overlapping strata [5, 7]. For SS, clusters need to be sampled from each stratum. This leaves us with the problem of deciding how to allocate samples. To this end, we adopt proportional allocation [7], which requires the sample size in each stratum to be proportional to the number of sampling units in that stratum. Therefore, relying on stratification, a Stratified TWCS (STWCS) design can be adopted where, in each stratum $q$, TWCS is applied with second stage size $m$ to obtain an unbiased estimator $\hat{\mu}_{q,w,m}$.

Let $E_q$ be the set of $N_q$ entities in the stratum $q$, $C_q = \{G[e] \mid e \in E_q\}$ the $q^{th}$ stratum cluster family, and $C_q = \sum_{i=1}^{N_q} M_i$ its cardinality. Then, denoting $W_q = C_q/M$ as the weight of the $q^{th}$ stratum, the unbiased estimator of $\mu(G)$ under STWCS is

$$\hat{\mu}_{ss} = \sum_{q=1}^{Q} W_q \cdot \hat{\mu}_{q,w,m} \tag{4}$$

with estimation variance $V(\hat{\mu}_{ss}) = \sum_{q=1}^{Q} W_q^2 V(\hat{\mu}_{q,w,m})$.

## 4 INTERVAL ESTIMATION

To quantify the uncertainties in the sampling procedure, we need to estimate the CI on the considered sample. A CI tells us that at a given level of certainty $1-\alpha$, if the underlying model is correct, the true value in the population will likely be in the identified range. The larger the CI, the more uncertain the observation. No unique solution exists to compute CIs for a given estimator. However, recalling that KG accuracy can be defined as the proportion of correct triples ($\tau$) over the total number of triples ($M$) in the KG, binomial CIs can be considered [4]. Several methods exist to construct binomial CIs [29, 36]. Nevertheless, the state-of-the-art KG accuracy estimation methods [13, 32] all adopt the standard normal approximation interval, also known as the Wald interval [5].

We first highlight the limitations of the Wald method for the task of efficient KG accuracy evaluation. Then, we present a family of binomial methods that overcome Wald limitations, providing reliable CIs for KG accuracy estimation. The methods are the Wilson interval [42], its continuity-corrected version [29], and the Agresti-Coull interval [1]. Finally, we conduct a theoretical comparison between the CIs, which we use to identify the method providing the best trade-off between efficiency and reliability – namely, Wilson.

### 4.1 The Wald Interval

The Wald interval relies on normal approximation and is obtained by inverting the acceptance region of the Wald large-sample normal test [5]:

$$\left| \frac{\hat{\mu} - \mu}{\sqrt{V(\hat{\mu})}} \right| \leq z_{\alpha/2} \tag{5}$$

where $\mu$ and $\hat{\mu}$ represent the true and estimated KG accuracies, $V(\hat{\mu})$ the estimation variance, and $z_{\alpha/2}$ the critical value of the standard normal distribution for a given significance level $\alpha$. Normal approximation requires that the point estimator $\hat{\mu}$ takes the form of the mean of $n_{\mathcal{S}}$ independent and identically distributed (i.i.d.) random variables with expectation equal to $\mu$. If the sample size $n_{\mathcal{S}}$ is sufficiently large,[1] then, by the Central Limit Theorem (CLT) [5], the $1-\alpha$ CI of $\mu$ can be constructed as

$$\hat{\mu} \pm z_{\alpha/2} \sqrt{V(\hat{\mu})} \tag{6}$$

The larger the sample size, the more *continuous* the approximation, and the more confident we can be in $\hat{\mu}$, so the CI shrinks as $n_{\mathcal{S}}$

[1] The rule of thumb is $n_{\mathcal{S}} \geq 30$ [17].

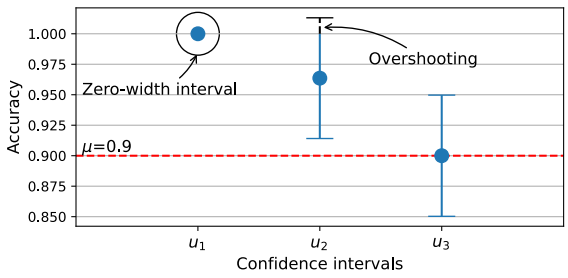


**Figure 2: Limitations of the Wald intervals: zero-width interval (circle) and overshooting (dashed line). CIs for the three users in Example 1, computed with 95% (i.e., $\alpha = 0.05$) Wald intervals on a KG with (true) accuracy $\mu = 0.9$.**

increases. However, although the Wald interval is appealing due to its simplicity, it is fundamentally flawed [4, 38].

Let us consider the case where $n_S$ is small or $\hat{\mu}$ is close to zero or one. In this context, two issues arise. First, when $\hat{\mu}$ tends to zero or one, the estimation variance $V(\hat{\mu})$ tends to zero, and the interval narrows to zero width. A **zero-width interval** implies certainty, leading to underestimating the error. Second, for $\hat{\mu} < z_{\alpha/2}\sqrt{V(\hat{\mu})}$ or $\hat{\mu} > 1 - z_{\alpha/2}\sqrt{V(\hat{\mu})}$, the interval **overshoots** the $[0, 1]$ boundaries. However, unlike the normal distribution – which is unconstrained – $\hat{\mu}$ cannot exceed the range $[0, 1]$, and thus the approximation fails.

**Example 1.** Let us assume that three independent users $u_1, u_2, u_3$ want to estimate the accuracy of a KG with (unknown) $\mu = 0.9$. Each one employs the evaluation framework described in Section 2.3, adopting SRS as a sampling strategy, with a significance level $\alpha = 0.05$, and requiring a MoE below $\varepsilon = 0.05$. Let us also assume that the evaluation procedure halted – i.e., MoE $\leq \varepsilon$ – with the following sample sizes and corresponding estimated accuracies: [2]

$u_1$: $n_1 = 30$, $\hat{\mu}_1 = 1.0000$
$u_2$: $n_2 = 55$, $\hat{\mu}_2 = 0.9636$
$u_3$: $n_3 = 140$, $\hat{\mu}_3 = 0.9000$

Based on these values, we can compute the estimation variances: $V(\hat{\mu}_1) = 0.0000$, $V(\hat{\mu}_2) = 0.0006$, and $V(\hat{\mu}_3) = 0.0006$. Then, by using Equation (6), we obtain the corresponding CIs: $\text{CI}_1 = [1.0000, 1.0000]$, $\text{CI}_2 = [0.9156, 1.0116]$, and $\text{CI}_3 = [0.8520, 0.9480]$. Figure 2 depicts the computed CIs. We can see that the evaluation procedure generated a zero-width interval for the user $u_1$. In other words, the procedure is certain that the estimated value $\hat{\mu}_1$ is correct, but we see that it leads to a 10% deviation from the true value $\mu$. As for the interval generated for user $u_2$, in addition to not containing the true value $\mu$, it also overshoots the upper limit – failing to approximate the underlying binomial distribution. The only user for whom the procedure works correctly is $u_3$, who obtains a reliable estimate of the true value accompanied by a proper interval. Hence, out of three situations, only one gets a reliable estimate, while the other two lead to misleading and erratic estimations.

[2]The values reported were cherry-picked for the example from real evaluations on a KG with $\mu = 0.9$. Therefore, they represent realistic situations.

Zero-width intervals further aggravate a well-known problem of the Wald interval: the erratic behavior of the **coverage probability** [4]. The coverage probability represents the probability that a CI contains the parameter of interest ($\mu$, in our case). For a given significance level $\alpha$, CIs are expected to have nominal coverage probability $1 - \alpha$. On the other hand, following a frequentist approach, the empirical coverage probability is defined as the proportion of trials where the CI surrounds the parameter. Ideally, we would like the empirical coverage to be the same as the nominal coverage. Yet, due to the discreteness and skewness of the underlying binomial distribution, the Wald interval often reaches empirical coverage far lower than nominal coverage [4]. This weakens the statistical guarantees, limiting the reliability of the KG accuracy evaluation procedure. We need better methods for building CIs, capable of handling small samples and skewed observations while reaching coverage probabilities consistently closer to the nominal value $1 - \alpha$.

## 4.2 The Wilson Interval

An alternative to the Wald interval is the Wilson interval [42], a binomial CI based on inverting the test in Equation (5) that uses the null standard error instead of the estimated standard error. The Wilson interval assumes the sample was obtained via SRS [42]. Based on this assumption, the test can thus be rewritten as:

$$\left| \sqrt{\frac{n_S}{\mu(1-\mu)}} \cdot (\hat{\mu} - \mu) \right| \leq z_{\alpha/2} \tag{7}$$

Then, by solving the algebraic inequality in $\mu$, we obtain

$$\frac{\hat{\mu} + \frac{z^2_{\alpha/2}}{2n_S}}{1 + \frac{z^2_{\alpha/2}}{n_S}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n_S}} \cdot \sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{n_S} + \frac{z^2_{\alpha/2}}{4n^2_S}} \tag{8}$$

which can be broken down into two parts: a relocated center estimate (left of the ± sign) and a corrected standard deviation (right of the ± sign). The Wilson interval has theoretical appeal, as it is the inversion of the CLT approximation to the family of equal tail tests of $H_0 : \mu = \mu_0$ [42]. Hence, the null hypothesis is accepted based on the CLT approximation if and only if $\mu_0$ is in this interval.

Reconsidering Example 1, we can use the data from $u_1$, $u_2$, and $u_3$ procedures to compute the Wilson interval. By applying Equation (8), we obtain $\text{CI}_1 = [1.0000, 0.8865]$, $\text{CI}_2 = [0.9900, 0.8767]$, $\text{CI}_3 = [0.9395, 0.8391]$.[3] As we can see, the zero-width interval ($\text{CI}_1$) now has a proper width, and the overshooting interval ($\text{CI}_2$) now falls within the $[0, 1]$ range. Also note that the Wilson deviation never reaches zero for values of $\hat{\mu}$ close to zero or one, as opposed to the Wald deviation that always converges to zero at boundaries, thus leading to zero-width intervals. Finally, compared to the Wald interval, the Wilson interval is *asymmetric*, as the center estimate is pushed towards the center of the (accuracy) range.

Although the Wilson interval improves over the Wald interval in several respects, its coverage probability still has some downward spikes when the accuracy is near the boundaries [4]. These spikes occur for all sample sizes $n_S$ and significance levels $\alpha$, but can be removed by using a one-sided Poisson approximation for $\tau_S$ close to 0 or $n_S$ [4]. That is, we can replace the lower bound of the Wilson

[3]Note that by using the Wilson interval none of the three users would obtain a MoE $\leq \varepsilon$ with that data and the iterative procedure would therefore have to continue.

interval with $\lambda_{\tau_S}/n_S$, for $\tau_S$ close to 0, and the upper bound with $1-\lambda_{\tau_S}/n_S$, for $\tau_S$ close to $n_S$. Using the relationship between the Poisson and $\chi^2$ distributions, the $\lambda_{\tau_S}$ can be formally expressed in terms of the $\chi^2$ quantiles [4]:

$$\lambda_{\tau_S} = \begin{cases} \frac{1}{2}\chi^2_{2\tau_S,\alpha} & \text{for } \tau_S \text{ close to 0,} \\ \frac{1}{2}\chi^2_{2(n_S-\tau_S),\alpha} & \text{for } \tau_S \text{ close to } n_S. \end{cases} \tag{9}$$

where $\chi^2_{2\tau_S,\alpha}$ and $\chi^2_{2(n_S-\tau_S),\alpha}$ denote the $100\alpha^{th}$ percentiles of the $\chi^2$ distribution with $2\tau_S$ and $2(n_S-\tau_S)$ degrees of freedom, respectively. Note that there is no strict rule for how close $\tau_S$ should be to 0 or $n_S$ to apply the correction; the choice is left to the user.

## 4.3 The Continuity-Corrected Wilson Interval

To further boost the coverage probability of the Wilson interval, a **continuity correction** mechanism can be applied [29]. The mechanism expands the interval by adding an extra $1/2n_S$ on either side of the center estimate. This correction term is sufficient to better adjust the boundaries of the interval to reflect the discrete nature of the underlying binomial distribution. Specifically, the continuity-corrected Wilson interval consists of all $\mu$ such that $|\hat{\mu}-\mu| - 1/2n_S \leq z_{\alpha/2}\sqrt{\mu(1-\mu)/n_S}$, leading to the following closed form expressions for the lower ($L$) and upper ($U$) bounds:

$$L = \frac{\hat{\mu} + \frac{z^2_{\alpha/2}}{2n_S} - \frac{\frac{1}{2n_S} + \frac{z^2_{\alpha/2}}{n_S}\sqrt{\frac{z^2_{\alpha/2}}{4} - \frac{1}{2} - \frac{1}{4n_S} + \hat{\mu}\{n_S(1-\hat{\mu})+1\}}}{1+\frac{z^2_{\alpha/2}}{n_S}}}{1+\frac{z^2_{\alpha/2}}{n_S}}$$

$$U = \frac{\hat{\mu} + \frac{z^2_{\alpha/2}}{2n_S} + \frac{\frac{1}{2n_S} + \frac{z^2_{\alpha/2}}{n_S}\sqrt{\frac{z^2_{\alpha/2}}{4} + \frac{1}{2} - \frac{1}{4n_S} + \hat{\mu}\{n_S(1-\hat{\mu})-1\}}}{1+\frac{z^2_{\alpha/2}}{n_S}}}{1+\frac{z^2_{\alpha/2}}{n_S}} \tag{10}$$

However, the continuity correction reintroduces overshooting as $\hat{\mu}$ approaches 0 or 1. To prevent this, it is necessary to employ min and max constraints to ensure that $L \in [0, \hat{\mu}]$ and $U \in [\hat{\mu}, 1]$.

**Proposition 1.** By construction, the continuity-corrected Wilson interval always contains the (uncorrected) Wilson interval.

## 4.4 The Agresti-Coull Interval

The Agresti-Coull interval [1] combines the Wald interval's simplicity with the Wilson interval's reliability. Specifically, the method adopts the familiar form presented in Equation (6), but replaces the center estimate of the Wald interval with that of the Wilson interval. Denoting $\tilde{\tau} = \tau_S + z^2_{\alpha/2}/2$ and $\tilde{n} = n_S + z^2_{\alpha/2}$, the Wilson center estimate can be rewritten as $\tilde{\mu} = \tilde{\tau}/\tilde{n}$, leading to the interval

$$\tilde{\mu} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\mu}(1-\tilde{\mu})}{\tilde{n}}} \tag{11}$$

**Lemma 1.** Agresti-Coull CIs are never shorter than Wilson CIs.

Proof. Since the intervals share the same center estimate, we only need to prove that

$$z_{\alpha/2}\sqrt{\frac{\tilde{\mu}(1-\tilde{\mu})}{\tilde{n}}} \geq \frac{z_{\alpha/2}}{1+\frac{z^2_{\alpha/2}}{n_S}} \cdot \sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{n_S} + \frac{z^2_{\alpha/2}}{4n^2_S}}$$

Recalling that $\tilde{\mu} = \tilde{\tau}/\tilde{n}$, where $\tilde{\tau} = \tau_S + z^2_{\alpha/2}/2$ and $\tilde{n} = n_S + z^2_{\alpha/2}$, and after some algebraic passages, we obtain $(n_S - 2\tau_S)^2 \geq 0$ which proves the theorem. Notably, the two intervals have the same width only when $\tau_S = n_S/2$; that is, when $\hat{\mu} = 0.5$. □

## 4.5 Interval Comparison

The minimization problem, outlined in Section 2.2, hinges on the convergence of the CI width, enforced by the constraint MoE $\leq \varepsilon$. As a result, to enhance the convergence rate and thus reduce the number of annotations required to evaluate the KG accuracy, we prefer methods that build small CIs. At the same time, we also seek reliable intervals that consistently achieve coverage probabilities closer to the nominal value $1-\alpha$. However, when $\hat{\mu}$ approaches zero or one, reliable CIs tend to be larger than unreliable ones, as shown by the comparison between Wald and Wilson in Example 1. Therefore, we compare the considered CIs to investigate which interval presents the best trade-off between efficiency and reliability.

To conduct this comparison, we evaluate both the expected width and the average expected width of the intervals. The expected width is computed as follows:

$$E_{n_S,\mu}(\text{width(CI)}) = \sum_{\tau_S=0}^{n_S} (U(\tau_S, n_S) - L(\tau_S, n_S)) \binom{n_S}{\tau_S} \mu^{\tau_S}(1-\mu)^{n_S-\tau_S} \tag{12}$$

where $U$ and $L$ represent the upper and lower bounds of the CI, respectively. For the calculation of the average expected width, we integrate the expected width over the accuracy interval $[0, 1]$, as expressed by $\int_0^1 E_{n_S,\mu}(\text{width(CI)})d\mu$.

Figure 3 shows the expected width of the four considered CIs for $n_S = 30$ and $\alpha = 0.05$. Notably, the Wald interval is the shortest when $\mu \leq 0.137$ or $\mu \geq 0.863$, whereas Wilson becomes the shortest when $0.137 < \mu < 0.863$. The small-to-zero width of the Wald interval as $\mu$ approaches zero or one underscores its unreliability in these situations. Conversely, the method exhibits large widths when accuracy moves towards the center of the range, making it less efficient compared to both Wilson and Agresti-Coull. This insight emphasizes that not only does the Wald method yield unreliable CIs near zero or one, but it is also less efficient than Wilson and Agresti-Coull when accuracy deviates from boundaries and normal approximation proves more effective. As expected, the continuity-corrected Wilson is always larger than the uncorrected Wilson, and Agresti-Coull equals Wilson only when $\mu = 0.5$ (cf. Lemma 1). Hence, the Wilson method emerges as the best trade-off between efficiency and reliability among the considered CIs.

This conclusion gains further support when we examine the average expected width of the considered CIs, shown in Figure 4 for sample sizes ranging from 30 to 50 with $\alpha = 0.05$. The plot shows that Wald and Wilson intervals exhibit nearly identical average expected widths. This indicates that Wilson compensates for the larger CIs obtained when accuracy is close to zero or one by yielding shorter intervals as accuracy moves towards the center of the range. In contrast, both Agresti-Coull and continuity-corrected Wilson are notably larger than Wilson.

On a side note, it is worth mentioning that as the sample size $n_S$ increases, the CIs shrink, and the average expected differences

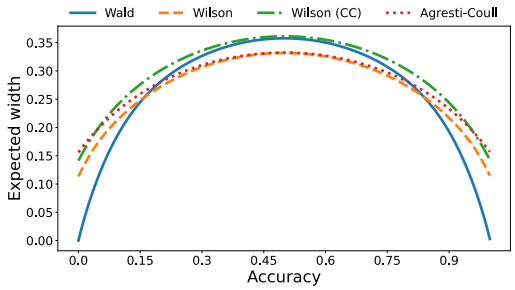


**Figure 3: Expected width of Wald, Wilson, continuity-corrected Wilson (CC), and Agresti-Coull CIs for $n_S = 30$ and $\alpha = 0.05$.**

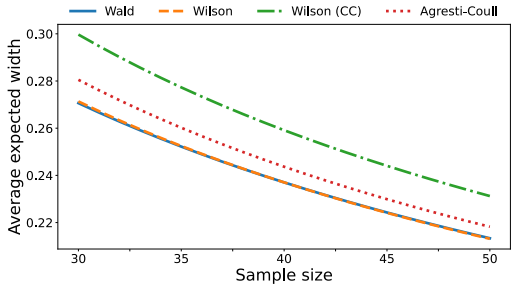


**Figure 4: Avg. expected width of Wald, Wilson, continuity-corrected Wilson (CC), and Agresti-Coull CIs for $n_S \in [30, 50]$ and $\alpha = 0.05$.**

between them diminish. Nevertheless, alongside Wald, Wilson remains the most efficient among the intervals while steering away from the latter's limitations. In Section 7.1, we empirically validate the above observations on real-world KGs.

The Wilson method builds tight intervals when accuracy is far from boundaries and becomes more lenient as accuracy converges towards zero or one, emerging as the best trade-off between efficiency and reliability among the considered CIs.

# 5 DESIGN EFFECT ADJUSTMENT

Wilson and the other binomial intervals are constructed assuming the use of SRS to obtain the sample. Therefore, they cannot manage effects such as clustering and stratification. As a consequence, using them on complex sampling designs like TWCS and STWCS would violate the assumption of a SRS design, leading to unstable CIs [22] (see Section 7.3). To adjust these intervals for cluster and stratified sampling, the **design effect** should be used [21, 22].

The design effect denotes the ratio of the estimation variance obtained via a complex design $\mathcal{S}$ to the estimation variance based on a SRS sample of the same size: $\text{Deff} = V(\hat{\mu}_S)/V(\hat{\mu}_s)$. In other words, the design effect indicates inflation ($\text{Deff} > 1$) or deflation ($\text{Deff} < 1$) in the variance of an estimator for a given parameter of interest when using a complex design instead of SRS. The design effect can be used to compute the **effective sample size** [21], which is defined as $n_{\text{eff}} = n_S/\text{Deff}$. The effective sample size represents the sample size required under a SRS design to obtain the same CI width achieved under the complex design. Thus, the effective sample size can be used to adjust binomial CIs to complex sampling designs by replacing $n_S$ with $n_{\text{eff}}$ in Equations (8)-(11).

Different design effects can be obtained depending on the components used in the sampling design. We provide the design effects for the cluster and stratified samplings, also proposing an adaptive strategy to compute the effective sample size for stratified sampling, which depends on the boundary conditions.

## 5.1 Clustering Effect

The combination of WCS in the first stage and SRS in the second stage makes TWCS an Equal Probability of Selection Method (EPSEM) design. Hence, we can apply the ultimate cluster sample approximation [20] reducing the variance of TWCS to $V(\hat{\mu}_{w,m}) = V(\hat{\mu}_s)\{1 + (\bar{m} - 1)\rho\}$, where $\bar{m}$ is the average second stage size and $\rho$ the Intracluster Correlation Coefficient (ICC) [7]. The ICC measures the degree of correlation among triples within the same cluster. To estimate ICC from the considered sample, we adopt the unbiased formula proposed by Fisher [12]:

$$\rho = \frac{\bar{m}}{\bar{m} - 1} \cdot \frac{\frac{1}{(n-1)} \cdot \sum_{i=1}^{n} (\hat{\mu}_i - \mu_{n_S})^2}{\frac{\sum_{i=1}^{n} \sum_{j=1}^{\min\{M_i, m\}} (\mathbb{1}_T(t_{ij}) - \mu_{n_S})^2}{\sum_{i=1}^{n} \min\{M_i, m\} - 1}} - \frac{1}{\bar{m} - 1} \tag{13}$$

where $\mu_{n_S}$ is the sample accuracy, defined as $\frac{\sum_{i=1}^{n} \sum_{j=1}^{\min\{M_i,m\}} \mathbb{1}_T(t_{ij})}{\sum_{i=1}^{n} \min\{M_i,m\}}$.

Thus, the design effect for TWCS becomes:

$$\text{Deff} = \frac{V(\hat{\mu}_{w,m})}{V(\hat{\mu}_s)} = \frac{V(\hat{\mu}_s)\{1 + (\bar{m} - 1)\rho\}}{V(\hat{\mu}_s)} = 1 + (\bar{m} - 1)\rho \tag{14}$$

Note that using the unbiased formula to estimate ICC allows the design effect to take values both greater and lower than one. In other words, no a priori assumption is made about the efficiency of TWCS with respect to SRS.

## 5.2 Stratification Effect

The STWCS design combines stratification and clustering effects. The use of proportional allocation makes STWCS an EPSEM design [7]. Therefore, we can apply the ultimate cluster sample approximation [20] and reduce the STWCS variance to

$$V(\hat{\mu}_{ss}) = \sum_{q=1}^{Q} W_q^2 V(\hat{\mu}_q)\{1 + (\bar{m}_q - 1)\rho_q\} \tag{15}$$

where $V(\hat{\mu}_q)$ represents the estimation variance of the $q^{th}$ stratum under SRS, while $\bar{m}_q$ and $\rho_q$ are the corresponding average second stage size and ICC, respectively. Based on Equation (15), the design effect can be defined as

$$\text{Deff} = \frac{\sum_{q=1}^{Q} W_q^2 V(\hat{\mu}_q)\{1 + (\bar{m}_q - 1)\rho_q\}}{V(\hat{\mu}_s)} \tag{16}$$

The design effect can be further decomposed under some conditions.

**Lemma 2.** If the strata (estimated) variances $\hat{\sigma}_q^2$ are approximately the same, the design effect can be expressed as

$$\text{Deff} = \sum_{q=1}^{Q} W_q^2 \left(\frac{n_S}{n_q}\right) \text{Deff}_q \tag{17}$$

Proof. Let us assume that strata variances are approximately the same, that is, $\hat{\sigma}_q^2 \approx \hat{\sigma}^2$. Under this assumption, the variance of

STWCS in Equation (15) becomes

$$V(\hat{\mu}_{ss}) = \hat{\sigma}^2 \sum_{q=1}^{Q} \frac{W_q^2}{n_q}\{1 + (\bar{m}_q - 1)\rho_q\}$$

Dividing $V(\hat{\mu}_{ss})$ by $V(\hat{\mu}_s)$, we obtain

$$\text{Deff} = \frac{\hat{\sigma}^2 \sum_{q=1}^{Q} \frac{W_q^2}{n_q}\{1 + (\bar{m}_q - 1)\rho_q\}}{\frac{\hat{\sigma}^2}{n_S}}$$

$$= \sum_{q=1}^{Q} W_q^2 \left(\frac{n_S}{n_q}\right)\{1 + (\bar{m}_q - 1)\rho_q\} = \sum_{q=1}^{Q} W_q^2 \left(\frac{n_S}{n_q}\right)\text{Deff}_q$$

□

**Remark 1.** With Deff as in Equation (17), $\text{Deff}_q$ represents the $q^{th}$ stratum design effect from clustering effects.

**Lemma 3.** If the strata (estimated) ICCs $\rho_q$ are approximately the same, the design effect can be expressed as

$$\text{Deff} = \{1 + (\bar{m} - 1)\rho\}\,\text{Deff}_q \tag{18}$$

Proof. Let us assume that strata ICCs are approximately the same, that is, $\rho_q \approx \rho$. Also consider that, under STWCS, $\bar{m}_q$ can be safely approximated by $\bar{m}$, i.e. $\bar{m}_q \approx \bar{m}$. Then, the design effect in Equation (16) becomes

$$\text{Deff} = \{1 + (\bar{m} - 1)\rho\} \cdot \frac{\sum_{q=1}^{Q} W_q^2 V(\hat{\mu}_q)}{V(\hat{\mu}_s)} = \{1 + (\bar{m} - 1)\rho\} \cdot \text{Deff}_q$$

□

**Remark 2.** With Deff as in Equation (18), $\text{Deff}_q$ represents the $q^{th}$ stratum design effect from element stratification.

**Lemma 4.** If both strata (estimated) variances and ICCs are approximately the same, the design effect can be expressed as

$$\text{Deff} = 1 + (\bar{m} - 1)\rho \tag{19}$$

Proof. Let us assume that $\hat{\sigma}_q^2 \approx \hat{\sigma}^2$ and $\rho_q \approx \rho$. Considering that $\bar{m}_q \approx \bar{m}$ under STWCS, the design effect in Equation (16) becomes

$$\text{Deff} = \{1 + (\bar{m} - 1)\rho\} \cdot \sum_{q=1}^{Q} W_q^2 \left(\frac{n_S}{n_q}\right)$$

With proportional allocation, $n_q = \frac{n_S C_q}{M}$ [7]. This reduces the design effect to

$$\text{Deff} = \{1 + (\bar{m} - 1)\rho\} \cdot \sum_{q=1}^{Q} W_q^2 \left(\frac{M}{C_q}\right)$$

$$= \{1 + (\bar{m} - 1)\rho\} \cdot \sum_{q=1}^{Q} W_q = 1 + (\bar{m} - 1)\rho$$

□

**Remark 3.** Under Equation (19), the design effect for STWCS takes the same form of the (unstratified) TWCS design effect.

Once the design effect is computed via one of the Equations (16)-(19), depending on which conditions are met, the effective sample size can be derived and used to adjust binomial CIs. To further counteract the undercoverage introduced by stratified clustering designs, Korn and Graubard [23] suggest multypling the effective sample size by a design factor, later refined by Dean and Pagano [9]:

$$n_{\text{eff}}^* = n_{\text{eff}} \cdot \left(\frac{z_{\alpha/2}}{t_{n-Q,\alpha/2}}\right)^2 \tag{20}$$

where $t_{n-Q,\alpha/2}$ denotes the $100(\alpha/2)$th percentile of the Student's $t$-distribution with $n - Q$ degrees of freedom. When the number of clusters ($n$) in the sample is relatively small, as it generally happens in practical cases, the design factor is less than one. Consequently, the effective sample size is reduced, resulting in wider intervals and mitigating, to a degree, undercoverage behaviors.

Thus, to compute the effective sample size for STWCS, we present a procedure based on adaptive design effect adjustment, reported in Algorithm 1. Given an STWCS stratified sample, we first derive the strata variances and ICCs (lines 1-2). Then, we compute the pairwise differences (lines 3-9) and check whether these differences are smaller than a given threshold $\beta_1$, for variances, and $\beta_2$, for ICCs (lines 10,12,14). Depending on which conditions are met, we apply one of the Equations (16)-(19) to obtain the design effect (lines 11,13,15,17). From this, we derive the effective sample size, which is finally corrected by the design factor (lines 19-20).

The proposed design effect adjustments enable binomial CIs – originally designed under the assumption of SRS – to account for clustering and stratification effects. The adaptive strategy proposed for stratification facilitates a smoother derivation of the design effect based on the boundary conditions.

**Algorithm 1** Adaptive Design Effect Adjustment Procedure for Effective Sample Size

**Input:**
A KG $G$;
The user-required significance level $\alpha$;
The $\beta_1$ and $\beta_2$ thresholds;
An STWCS sample $G_S = \{G_1, \ldots, G_Q\}$, where $G_q$ is the $q$th stratum TWCS sample.

**Output:** The effective sample size $n_{\text{eff}}^*$ for the stratified sample $G_S$.

1: Compute strata sample variances: $\hat{\sigma}_q^2 = \hat{\mu}_q(1 - \hat{\mu}_q)$
2: Compute strata ICCs $\rho_q$ using Eq(13)
3: Initialization: $\Delta_{\hat{\sigma}^2} \leftarrow$ empty list; $\Delta_\rho \leftarrow$ empty list
4: **for** $i = 1$ **to** $Q - 1$ **do**
5: **for** $j = i + 1$ **to** $Q$ **do**
6: $\Delta_{\hat{\sigma}^2} \leftarrow \text{diff}(\hat{\sigma}_i^2, \hat{\sigma}_j^2)$
7: $\Delta_\rho \leftarrow \text{diff}(\rho_i, \rho_j)$
8: **end for**
9: **end for**
10: **if** all($\delta_{\hat{\sigma}^2} < \beta_1$ **for** $\delta_{\hat{\sigma}^2}$ **in** $\Delta_{\hat{\sigma}^2}$) **and** all($\delta_\rho < \beta_2$ **for** $\delta_\rho$ **in** $\Delta_\rho$) **then**
11: Calculate Deff using Eq(19)
12: **else if** all($\delta_{\hat{\sigma}^2} < \beta_1$ **for** $\delta_{\hat{\sigma}^2}$ **in** $\Delta_{\hat{\sigma}^2}$) **then**
13: Calculate Deff using Eq(17)
14: **else if** all($\delta_\rho < \beta_2$ **for** $\delta_\rho$ **in** $\Delta_\rho$) **then**
15: Calculate Deff using Eq(18)
16: **else**
17: Calculate Deff using Eq(16)
18: **end if**
19: Compute effective sample size: $n_{\text{eff}} = n_S/\text{Deff}$
20: Multiply by design factor: $n_{\text{eff}}^* = n_{\text{eff}} \cdot \left(\frac{z_{\alpha/2}}{t_{n-Q,\alpha/2}}\right)^2$
21: **return** $n_{\text{eff}}^*$

**Table 2: Data statistics for YAGO, NELL, DisGeNET, and SYN 100M.**

| | YAGO | NELL | DisGeNET | SYN 100M |
|---|---|---|---|---|
| Number of facts | 1,386 | 1,860 | 2,999,087 | 101,415,011 |
| Number of clusters | 822 | 817 | 21,243 | 5,000,000 |
| Average cluster size | 1.69 | 2.28 | 141.18 | 20.28 |
| Accuracy ($\mu$) | 0.99 | 0.91 | n/a | n/a |

# 6 EXPERIMENTAL SETUP

This section comprises five parts: (i) data presentation and selection rationale; (ii) synthetic label generation for the DisGeNET and SYN 100M datasets; (iii) annotation cost function for manual fact evaluation; (iv) implementation, parameter tuning, and computing resources; and, (v) evaluation procedure and selected metrics.

## 6.1 Datasets

Table 2 reports statistics for the considered KGs. YAGO and NELL [30] are adopted because they are publicly available reference KGs that have been used to evaluate the state-of-the-art methods for KG accuracy estimation in the literature [13, 30, 32]. [4]

**YAGO** is a sample drawn by Ojha and Talukdar [30] from the YAGO2 KG [16] – a large KG with general knowledge about people, cities, countries, movies, and organizations. The sample contains manually annotated accuracy labels for each fact. The ground-truth accuracy of YAGO is $\mu = 0.99$.

**NELL** is a sample drawn by Ojha and Talukdar [30] from the NELL KG [27]. The sample contains sports-related facts, mostly about athletes, coaches, teams, stadiums, etc. As with YAGO, manually annotated accuracy labels are available. The ground-truth accuracy of NELL is $\mu = 0.91$.

Moreover, we consider data from **DisGeNET** [33], one of the largest collections of facts about gene-disease associations, integrating data from expert-curated repositories, GWAS catalogs, animal models, and scientific literature [33]. DisGeNET is a highly specialized resource containing nearly three million factual triples, providing confidence scores for each fact. However, there are no available human-annotated accuracy labels. Due to the high level of expertise required to audit DisGeNET facts, manually evaluating its accuracy is not affordable. Therefore, we generate synthetic labels for DisGeNET to compare the different methods in-depth. To this end, we adopt two labeling schemes, presented in Section 6.2, that allow us to test the proposed methods under different conditions.

To test scalability, we also generate **SYN 100M**, a synthetic KG with over 100 million triples. Entity clusters were generated with a mean size of 20 and a standard deviation of 15. With SYN 100M we test the scalability of the proposed methods. We want to verify whether the findings obtained in YAGO, NELL, and DisGeNET KGs also hold when the KG size scales up. Again, we use synthetic labels to annotate the KG.

Hence, to evaluate the performance of the methods from different angles, we resort to two real datasets with real labels, YAGO and NELL, one real dataset with synthetic labels, DisGeNET, and one synthetic dataset with synthetic labels, SYN 100M.

[4] Gao et al. [13] evaluate also the MOVIE KG. The dataset was generated at Amazon and is not publicly available (personal communication). Qi et al. [32] used the public OPIEC KG [15]. However, the data subset employed and the annotations are unavailable.

## 6.2 Synthetic Label Generation

We consider two synthetic label generation models presented in [13].

**Triple Error Model (TEM):** the probability that a triple in the KG is correct is governed by a fixed error rate $\varepsilon_T \in [0, 1]$. This leads to a uniform distribution of correct labels across entity clusters in the KG. For our experiments, we vary $\varepsilon_T$ from 0.1 to 0.9.

**Cluster Error Model (CEM):** the number of correct triples in the $i^{th}$ entity cluster follows a binomial distribution $\mathbb{1}_E(G[e]) \sim B(M_i, p_i)$, where $p_i$ is defined by a sigmoid-like function:

$$p_i = \begin{cases} 0.5 + \varepsilon_{\mathcal{N}} & \text{if } M_i < k \\ \frac{1}{1+e^{-c(M_i-k)}} + \varepsilon_{\mathcal{N}} & \text{if } M_i \geq k \end{cases} \tag{21}$$

where, $\varepsilon_{\mathcal{N}}$ is an error term from a normal distribution with mean 0 and standard deviation $\sigma$, while $c \geq 0$ scales the influence of cluster size on entity accuracy. Together, $\varepsilon_{\mathcal{N}}$ and $c$ control the correlation between $M_i$ and $p_i$. This leads to a distribution of correct labels across entity clusters that depends, more or less strongly, on the size of clusters. For our experiments, we set $k = 3$, $\sigma = 0.01$ (in $\varepsilon_{\mathcal{N}}$), and $c = 0.001$, obtaining a KG sensitive to cluster characteristics.

Synthetic label generation facilitates experimentation and comparison of various sampling and evaluation methods, allowing for the assumption of diverse data characteristics on KGs. This ensures exploration of both efficiency and reliability when manual annotations are not available – or not affordable. In this regard, TEM and CEM can experiment under different accuracy distributions, enabling a thorough investigation into the performance of the considered methods across a broad spectrum of scenarios.

## 6.3 Cost Function

To measure the cost of manually evaluating the correctness of facts within the considered sample $G_S$, we adopt the cost function proposed in [13]. Based on the assumption that annotating a new fact for an entity that has been already identified reduces the annotation cost compared to assessing a new fact from unseen entities, the function is defined as follows:

$$\text{cost}(G_S) = |E_S| \cdot c_1 + |T_S| \cdot c_2 \tag{22}$$

where $c_1$ and $c_2$ are the average cost (in seconds) of entity identification and fact verification, respectively. For $c_1$ and $c_2$, we resort to the values estimated in [13], that is $c_1 = 45$ and $c_2 = 25$ (seconds).

## 6.4 Implementation

**Sampling Design:** we consider SRS (Section 3.1), TWCS (Section 3.2), and STWCS (Section 3.3) as sampling strategies. For TWCS and STWCS, Gao et al. [13] suggest setting the second stage size $m$ in the $\{3, 5\}$ range. Therefore, we set $m = 3$ for YAGO and NELL due to their small average cluster size. Instead, for DisGeNET and SYN, where the average cluster size is larger, we set $m = 5$. For STWCS, we stratify entity clusters by the degree centrality of the subject entity using the Cumulative Square Root of Frequency (Cumulative $\sqrt{F}$) [8]. For each KG, we consider a number of strata $Q = 2$.

**Interval Estimation:** to build CIs, we consider the Wald interval (Section 4.1) and compare its performance with our solution based on the Wilson interval (Section 4.2). We also evaluate Wilson performance against its continuity-corrected version (Section 4.3) and the Agresti-Coull interval (Section 4.4). For the Wilson interval,

we need to decide how close $\tau_{\mathcal{S}}$ should be to 0 or $n_{\mathcal{S}}$ to apply the one-sided Poisson correction. Accordingly to Brown et al. [4], we require $\tau_{\mathcal{S}} \in \{1, 2\}$ when $n_{\mathcal{S}} \leq 50$ and $\tau_{\mathcal{S}} \in \{1, 2, 3\}$ when $n_{\mathcal{S}} \geq 51$, to correct the lower bound, while $\tau_{\mathcal{S}} \in \{n_{\mathcal{S}} - 1, n_{\mathcal{S}} - 2\}$ when $n_{\mathcal{S}} \leq 50$ and $\tau_{\mathcal{S}} \in \{n_{\mathcal{S}} - 1, n_{\mathcal{S}} - 2, n_{\mathcal{S}} - 3\}$ when $n_{\mathcal{S}} \geq 51$, to correct the upper bound.

**Design Effect Adjustment:** we apply the design effect adjustment to the Wilson interval when using TWCS (Section 5.1) and STWCS (Section 5.2) schemes. For STWCS, we set the thresholds in the adaptive design effect adjustment procedure (Algorithm 1) to $\beta_1 = \beta_2 = 0.01$. To validate the benefits of applying design effect adjustments to the Wilson interval, we also present solutions that use Wilson without them. These versions are referred to as *vanilla*.

**Methods:** the configuration based on the Wald interval is the state-of-the-art adopted by efficient KG accuracy estimation solutions [13, 32]. We label these baselines as {method} (Wald). Instead, we label the newly proposed methods built upon the Wilson interval as {method} (Wilson).

**Computational Resources:** we implemented all methods in Python3 and performed all experiments on a Linux machine with an Intel Core i9-7980XE 2.60GHz processor and 64GB of memory.

## 6.5 Evaluation Procedure

We set the significance level $\alpha = 0.05$ and the upper bound for the MoE $\varepsilon = 0.05$. We set the minimum number of annotated triples to 30 and repeat the evaluation procedure 1,000 times for each method. The performance of the various methods is compared by the number of annotated triples, annotation cost (in hours), and empirical coverage. Results are reported once MoE $\leq 0.05$, ensuring the performance is compared when solutions satisfy the optimization constraint. For this reason, we do not include the width of CIs in the results – as all solutions will have MoE $\leq 0.05$. Likewise, we avoid reporting accuracy estimates, as all methods yield unbiased estimates with minimal difference ($\leq 0.02$) from ground-truth accuracy.

# 7 EXPERIMENTAL RESULTS

The four goals of the evaluation are to (i) compare the efficiency and reliability of Wald and binomial CIs; (ii) quantitatively analyze the limitations of the Wald method in building reliable CIs; (iii) investigate the performance of the proposed suite of methods against state-of-the-art solutions; and (iv) test their scalability.

## 7.1 Interval Comparison

We start by comparing the performance of Wald, Wilson, and the other binomial CIs to investigate which interval presents the best trade-off between efficiency and reliability. The evaluation involves assessing annotation costs and coverage probabilities across various KGs characterized by different levels of accuracy, sizes, and topologies. The considered KGs are YAGO ($\mu = 0.99$), NELL ($\mu = 0.91$), DisGeNET CEM ($\mu = 0.76$), and DisGeNET TEM with $\varepsilon_T = 0.5$ ($\mu = 0.5$). In this way, we can evaluate the performance of Wald, Wilson, and the other binomial CIs when accuracy is close to the boundaries (YAGO and NELL) and when it approaches the center of the range (DisGeNET CEM and TEM) – where normal approximation proves more effective. As a sampling strategy, we resort to SRS. This choice allows us to compare the capabilities of the considered CIs in their most natural form without introducing clustering and stratification effects. Such complexities can hamper the interpretation of the results, making it challenging to determine whether a particular outcome is attributable to the interval under consideration or other factors. Results are shown in Figure 5(a) for annotation costs (efficiency) and in Figure 5(b) for coverage probabilities (reliability).

***Efficiency.*** In Figure 5(a), we observe that the annotation costs increase for all solutions as we move towards KGs with accuracy deviating further from boundaries, reaching the highest costs when $\mu = 0.5$ (DisGeNET TEM). On the one hand, the Wald method obtains the lowest costs when accuracy is near the boundaries (i.e., YAGO and NELL), yielding performance gains up to 23%, 54%, and 66% compared to Wilson, its continuity-corrected version, and Agresti-Coull, respectively. Among the binomial intervals, Wilson emerges as the most efficient, having the smallest gap compared to Wald on YAGO (23% drop) and NELL (9% drop). On the other hand, top performance is obtained by Wilson and Agresti-Coull intervals as accuracy deviates from one (i.e., DisGeNET CEM and TEM). This corroborates the findings of the theoretical comparison in Section 4.5, where we demonstrated that Wilson is the most efficient among the considered binomial intervals while deviating from the unreliable widths obtained by the Wald method when accuracy approaches zero or one.

***Reliability.*** In Figure 5(b), we see that the Wald low annotation costs on YAGO and NELL come at the expense of reliability. The method achieves empirical coverages of 0.20 on YAGO and 0.82 on NELL. In both cases, the empirical coverages are far from the nominal coverage of 0.95 (see the dashed red line). Conversely, all binomial CIs exhibit coverage probabilities larger than nominal,[5] making the entire evaluation procedure up to four times more reliable. On DisGeNET CEM and TEM, instead, all methods achieve empirical coverage equal to or greater than nominal. This happens because the further we deviate from boundaries, the more the accuracy distribution resembles a normal distribution, thereby making the Wald method – based on normal approximation – more reliable.

> When accuracy approaches zero or one, Wilson emerges as the most efficient solution among the reliable ones. When accuracy deviates from boundaries, all solutions are reliable and Wilson is the most efficient overall. Hence, Wilson represents the best trade-off between efficiency and reliability.

## 7.2 Wald Limitations

We proceed by examining the limitations of the Wald interval, evaluating the extent to which overshooting and zero-width intervals affect KGs of different sizes, topologies, and levels of accuracy. To do so, we consider the same KGs as in Section 7.1 but also adopt TWCS and STWCS as sampling strategies. Results are shown in Figure 6. The number of iterations (over 1,000) affected by overshooting is in blue, whereas zero-width intervals are in red. We highlight the following observations.

As expected, the number of iterations where overshooting or zero-width intervals occur peaks when $\mu$ is near the boundaries

[5] When this happens, the obtained CIs are referred to as *conservative intervals*.

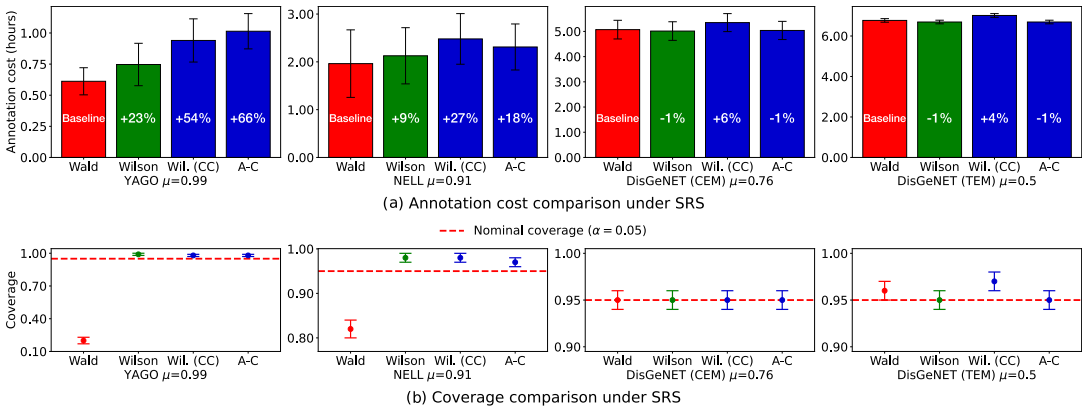


**Figure 5: Comparison of efficiency (a) and reliability (b) across Wald, Wilson, its continuity-corrected version, and the Agresti-Coull intervals under SRS on YAGO, NELL, DisGeNET (CEM), and DisGeNET (TEM) KGs. For efficiency (a), we also report the performance drop/gain (in %) obtained by binomial intervals compared to the Wald interval (baseline).**

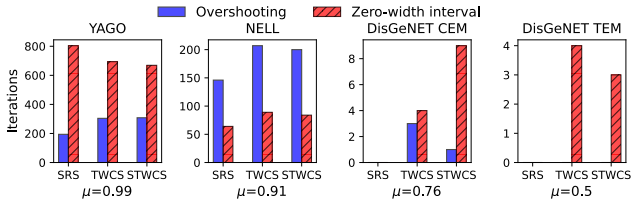


**Figure 6: Number of iterations (over 1,000) where the Wald interval overshoots or gets to zero-width for YAGO, NELL, DisGeNET CEM, and DisGeNET TEM KGs. We do not report numbers for the Wilson method as it does not produce overshooting or zero-width intervals.**

(i.e., YAGO and NELL). In YAGO, the number of iterations where the Wald method builds overshooting or zero-width intervals is close to – or even equals, in the case of SRS – the total number of 1,000 iterations. This leads to imprecise estimates, which affect the reliability of the considered estimators and make the efficiency aspect of the evaluation procedure a negligible feature.

On the other hand, the more the KG accuracy deviates from boundaries, the more the number of iterations affected by these problems decreases, reaching no more than nine iterations for DisGeNET CEM and four for DisGeNET TEM ($\varepsilon_T = 0.5$). However, even in these cases, the deviation from the true value can be significant, reaching up to 24% for DisGeNET CEM and 10% for DisGeNET TEM when zero-width intervals occur. As a result, users can incur estimations deemed error-free by the estimator but which are, in reality, far from the true accuracy of the KG.

None of the above limitations affect the Wilson interval.

> The Wilson method overcomes the limitations of Wald, preventing both overshooting and zero-width intervals, regardless of the considered sampling design and underlying KG.

## 7.3 Method Comparison

We continue with an in-depth comparison between proposed methods (Wilson) and baselines (Wald), focusing on state-of-the-art sampling strategies – namely, TWCS and STWCS. Additionally, to validate the benefits of applying the design effect to Wilson CIs under TWCS and STWCS, we also compare the proposed methods against versions that use Wilson without adjustments (Vanilla). The evaluation considers YAGO, NELL, DisGeNET CEM, and DisGeNET TEM ($\varepsilon_T = 0.5$) KGs. This evaluation explains how methods compare across various KGs, exhibiting different accuracy levels, sizes, and topologies. Results are depicted in Figure 7(a) for annotation costs and in Figure 7(b) for coverage probabilities. Moreover, we compare methods on DisGeNET TEM with $\varepsilon \in \{0.1, 0.5, 0.9\}$ to examine further the influence of the KG accuracy level on the performance of the different methods. Although synthetic, this experiment allows us to isolate the role of accuracy by fixing the underlying KG and only varying its accuracy distribution. Table 3 presents the results of this analysis, reporting the number of annotated triples and empirical coverage.

**Table 3: Performance on DisGeNET TEM with $\varepsilon_T \in \{0.1, 0.5, 0.9\}$.**

| | DisGeNET TEM | | | | | |
|---|---|---|---|---|---|---|
| | $\mu = 0.9$ ($\varepsilon_T = 0.1$) | | $\mu = 0.5$ ($\varepsilon_T = 0.5$) | | $\mu = 0.1$ ($\varepsilon_T = 0.9$) | |
| Method | Triples | Coverage | Triples | Coverage | Triples | Coverage |
| TWCS (Wald) | 118±50 | 0.82±0.02 | 379±67 | 0.94±0.02 | 119±48 | 0.83±0.02 |
| TWCS (Vanilla) | 134±36 | 0.93±0.01 | 382±2 | 0.95±0.01 | 136±35 | 0.93±0.01 |
| TWCS (Wilson) | 120±46 | 0.92±0.02 | 371±62 | 0.95±0.01 | 121±44 | 0.92±0.02 |
| STWCS (Wald) | 120±54 | 0.78±0.03 | 380±72 | 0.93±0.01 | 117±55 | 0.78±0.03 |
| STWCS (Vanilla) | 135±34 | 0.93±0.01 | 382±3 | 0.95±0.01 | 134±36 | 0.93±0.01 |
| STWCS (Wilson) | 125±51 | 0.91±0.02 | 376±74 | 0.93±0.02 | 125±50 | 0.92±0.02 |

***Wilson vs. Wald.*** Upon examining Figure 7, we observe a pattern akin to that seen under SRS (cf. Figure 5). Notably, on YAGO and NELL, Wilson solutions exhibit lower efficiency but higher reliability than Wald counterparts – making the evaluation procedure on YAGO up to two times more reliable. Besides, compared to the SRS scenario, the increase in annotation costs of proposed methods (Wilson) concerning baselines (Wald) is less pronounced. On NELL, STWCS (Wilson) even reduces annotation costs by 5% compared to STWCS (Wald) while attaining empirical coverage close to the nominal value (0.95), thereby increasing reliability by 26%. Conversely, on DisGeNET CEM and TEM ($\varepsilon_T = 0.5$), the performance of Wald

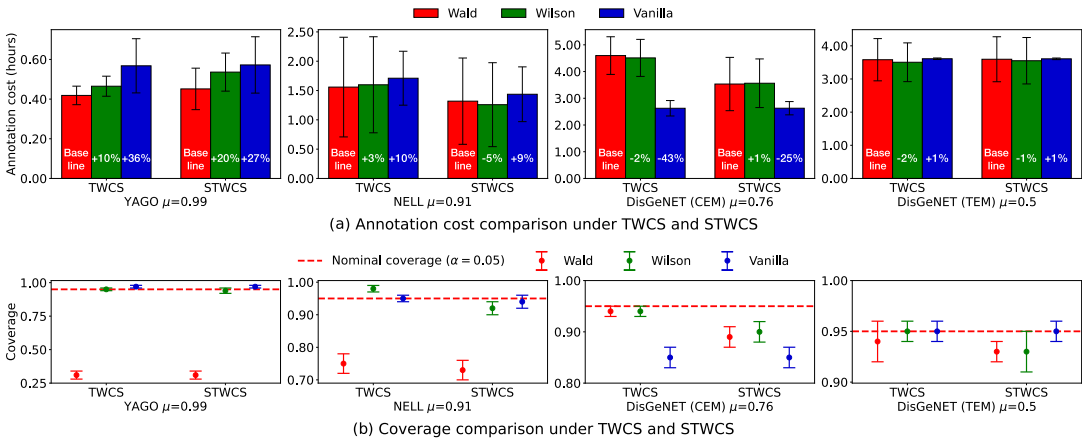


**Figure 7: Comparison of efficiency (a) and reliability (b) across Wald, Wilson, and its vanilla version under TWCS and STWCS on YAGO, NELL, DisGeNET (CEM), and DisGeNET (TEM) KGs. For efficiency (a), we also report the performance drop/gain (in %) obtained by adjusted and vanilla Wilson intervals compared to the Wald interval (baseline).**

and Wilson solutions becomes comparable, but Wilson solutions show higher efficiency.

The results presented in Table 3 provide additional evidence to support the above observations and confirm that the KG accuracy level significantly influences the method's performance. Indeed, the methods achieve the lowest annotation costs when KG accuracy is near the boundaries – whether they are zero ($\varepsilon_T = 0.9$) or one ($\varepsilon_T = 0.1$) – and incur the highest costs when KG accuracy is 0.5 ($\varepsilon_T = 0.5$), which is the accuracy level where the variance among triples' correctness (1 or 0) reaches its maximum.

Overall, the cost required to perform the evaluation procedure is notably lower with TWCS and STWCS compared to SRS (cf. Figure 5(a)), albeit at the expense of a reduction in coverage. This decrease can be attributed to the higher complexity of TWCS and STWCS, which, unlike SRS, involve clustering and stratification. Nevertheless, both TWCS and STWCS, when used in tandem with Wilson, achieve coverage probabilities close to the nominal value for all the considered KGs, unlike Wald solutions. Thus, Wilson again emerges as the best trade-off between efficiency and reliability.

***Wilson vs. Vanilla.*** Comparing Wilson and its vanilla version yields the following observations. Wilson demonstrates greater efficiency with TWCS and STWCS strategies on YAGO and NELL while maintaining comparable empirical coverage probabilities. Similar findings also emerge from the analysis of DisGeNET TEM KGs, where Wilson consistently requires fewer annotations than its vanilla counterpart, albeit with a negligible decrease in empirical coverage. In this regard, Agresti and Coull [1] argue that, for most applications, shorter intervals (i.e., more efficient) with empirical coverages lower than but close to the nominal confidence level are preferable to wider intervals with higher coverage (i.e., more reliable). Thus, design-effect-adjusted Wilson represents a preferable solution over vanilla Wilson.

A contrasting scenario emerges when examining DisGeNET CEM. Here, vanilla Wilson yields significantly lower annotation costs than the design-effect-adjusted Wilson for both TWCS and STWCS strategies. However, the cost-saving behavior of vanilla Wilson comes at the expense of notably lower empirical coverages compared to adjusted Wilson. Specifically, vanilla Wilson returns up to 150% more unreliable CIs for TWCS and 50% more for STWCS, with significant deviations from the true value – reaching up to 14% for both TWCS and STWCS. This outcome is expected given that vanilla Wilson is derived from SRS, and its application to TWCS and STWCS violates its underlying assumptions (see Section 5). Consequently, vanilla Wilson fails to account for clustering and stratification effects when used on DisGeNET CEM, where such effects are prominent. This results in too narrow intervals, thereby providing a false sense of precision. In contrast, design-effect-adjusted Wilson effectively overcomes this limitation, restoring empirical coverages to levels close to the nominal value.

> When combined with state-of-the-art sampling strategies, Wilson provides the best trade-off between efficiency and reliability. Furthermore, applying design effect adjustments proves essential to ensure stable CIs, especially when clustering and stratification effects are prominent.

## 7.4 Scalability

We investigate the scalability of the proposed methods by verifying whether the findings obtained in the previous experiments hold true when we consider KGs with similar accuracy levels but different sizes and topologies. To accomplish this, we compare methods on NELL, DisGeNET TEM ($\varepsilon_T = 0.1$), and SYN 100M ($\varepsilon_T = 0.1$). These KGs have similar accuracy levels ($\mu \approx 0.9$) but exhibit small (NELL), medium (DisGeNET), and large (SYN 100M) sizes with different topologies (cf. Table 2).

The experimental results are presented in Table 4, where we report the number of annotated triples and empirical coverage. We can highlight two key points. First, despite the increase in size – ranging across three orders of magnitude from NELL to DisGeNET, and two from DisGeNET to SYN 100M – the number of annotations and the empirical coverage remain consistent for all methods across the KGs. This indicates that while the evaluation procedure is influenced by the accuracy level of the underlying KG (cf. Table 3), it remains unaffected by its size or topology. This is

**Table 4: Performance on NELL (small), DisGeNET TEM (medium), and SYN 100M TEM (large) KGs.**

| | NELL | | DisGeNET TEM | | SYN 100M TEM | |
|---|---|---|---|---|---|---|
| | $\mu = 0.91$ | | $\mu = 0.9$ ($\varepsilon_T = 0.1$) | | $\mu = 0.9$ ($\varepsilon_T = 0.1$) | |
| Method | Triples | Coverage | Triples | Coverage | Triples | Coverage |
| SRS (Wald) | 107±40 | 0.82±0.02 | 125±43 | 0.83±0.02 | 122±43 | 0.83±0.02 |
| SRS (Wilson) | 116±33 | 0.98±0.01 | 132±35 | 0.94±0.02 | 131±34 | 0.93±0.02 |
| TWCS (Wald) | 124±68 | 0.75±0.03 | 118±50 | 0.82±0.02 | 122±52 | 0.83±0.02 |
| TWCS (Wilson) | 128±65 | 0.98±0.01 | 120±46 | 0.92±0.02 | 123±51 | 0.93±0.02 |
| STWCS (Wald) | 105±59 | 0.73±0.03 | 120±54 | 0.78±0.03 | 117±60 | 0.77±0.02 |
| STWCS (Wilson) | 100±57 | 0.92±0.02 | 125±51 | 0.91±0.02 | 122±57 | 0.88±0.02 |

an interesting outcome, as it not only shows that the procedure is insensitive to KG size – as also observed by Gao et al. [13] – but also to the topological structure of the considered KG.

Secondly, we observe consistent trends between baselines (Wald) and the proposed methods (Wilson) across different sizes and topologies. Specifically, proposed methods significantly improve coverage when the KG accuracy is near the boundaries while requiring a similar number of annotations. The same observations can also be found when comparing DisGeNET and SYN 100M across different values of $\varepsilon_T$. However, we do not report on these comparisons due to space reasons.

The performance of the proposed methods remains consistent across KGs with different sizes and topologies, maintaining high coverage probabilities at low annotation costs.

## 8 RELATED WORK

***Efficient KG Accuracy Evaluation.*** Ojha and Talukdar [30] were the first to recognize the need to efficiently estimate the accuracy of large-scale KGs, largely unexplored in prior work. To this end, they introduced KGEval, an iterative algorithm that alternates between two stages: control and inference mechanisms. Although pioneering, KGEval has two significant limitations. First, its probabilistic inference process can lead to the propagation of erroneous beliefs making it difficult to assess the bias introduced into the accuracy estimation. Secondly, the inference mechanism of KGEval does not scale to large-size KGs, as shown in [13]. For these reasons, we do not consider it in our work.

Following [30], Gao et al. [13] advocated the need for sampling strategies that generate representative samples of the KG. The authors pointed out that SRS fails to guarantee any relationship between sampled facts, thus incurring high annotation costs. To overcome this limitation, they resorted to cluster sampling strategies, identifying TWCS as the most suited for the task. However, their main focus was selecting an optimal sampling strategy to minimize annotation costs, neglecting the impact of CIs on the problem. For building CIs, the authors opted for the Wald method [5], known to underperform when used on binomial proportions [4, 38]. Our research addresses this limitation by (i) highlighting the drawbacks of Wald intervals and (ii) introducing and comparing binomial intervals. Among these, the Wilson interval [42] emerges as the best trade-off between efficiency and reliability. Furthermore, we introduce coverage probability as a metric to evaluate the reliability of KG accuracy estimation methods.

Gao et al. [13] also addressed the problem of evaluating an evolving KG. Since the construction of CIs is independent of this aspect, we leave the study of evolving KGs as future work.

Qi et al. [32] proposed an efficient human-machine collaborative framework to minimize annotation costs further. Inspired by [30], the proposed framework first creates inference graphs and then interleaves sampling and verification over them. The main focus of the work is on the verification aspect, which serves for the automatic inference of the veracity of additional facts. The sampling strategies, estimators, and CIs align with those considered in [13]. Since our research revolves around these specific aspects rather than the inference phase, we do not incorporate this method into our experiments. Note, however, that our contributions – being orthogonal to what was proposed by Qi et al. [32] – can be integrated into their approach to further boost efficiency while improving statistical reliability.

***Data Quality.*** Efficient KG accuracy evaluation lies under the umbrella of data quality, as accuracy represents one of the core data quality dimensions [40, 43]. Data quality encompasses several activities, such as data cleaning [6, 26, 39] and knowledge verification [24, 25], to name a few. Although related, these activities focus on specific aspects linked to KG quality and do not directly address efficient KG accuracy evaluation.

***Approximate Query Processing.*** Another related line of research is approximate query processing [14, 41], whose objective is to efficiently find an approximate answer as close as possible to the exact one. Hence, the devised sampling strategies focus on samples suited to approximate query results but not necessarily representative of the KG. Besides, efficient KG evaluation also involves optimizing a constrained minimization problem, thus introducing further constraints in both sampling strategies and estimators.

## 9 CONCLUSIONS

In this paper, we highlighted the problems of current state-of-the-art solutions for KG accuracy evaluation. Relying on the Wald method to build CIs, these approaches are hindered by zero-width and overshooting intervals, compromising the reliability of the estimations. To overcome these limitations, we introduced a family of binomial intervals, with Wilson being the most notable representative, and we adapted them to complex sampling designs, ensuring stable CIs. Through theoretical and empirical analyses, we identified Wilson as the best trade-off between efficiency and reliability.

Harnessing Wilson, we proposed solutions that advance the state-of-the-art. Extensive experiments across various real-life and synthetic KGs, characterized by different accuracy levels, sizes, and topologies, show that our solutions are (i) up to two times more reliable than state-of-the-art in cases where Wald intervals prove unreliable and (ii) more efficient when Wald intervals are reliable.

Following a thorough comparison of various state-of-the-art sampling strategies, including clustering (TWCS) and stratification (STWCS), we advise practitioners to adopt TWCS alongside Wilson intervals to assess KG accuracy in an efficient *and* reliable manner.

## ACKNOWLEDGMENTS

The work was supported by the HEREDITARY project, as part of the EU Horizon Europe program under Grant Agreement 101137074.

# REFERENCES


[1] A. Agresti and B. A. Coull. 1998. Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician* 52, 2 (1998), 119–126. https://doi.org/10.1080/00031305.1998.10480550

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007 (LNCS)*, Vol. 4825. Springer, 722–735. https://doi.org/10.1007/978-3-540-76298-0_52

[3] A. Bonifati, G. H. L. Fletcher, H. Voigt, and N. Yakovets. 2018. *Querying Graphs.* Morgan & Claypool Publishers. https://doi.org/10.2200/S00873ED1V01Y201808DTM051

[4] L. D. Brown, T. T. Cai, and A. DasGupta. 2001. Interval Estimation for a Binomial Proportion. *Statist. Sci.* 16, 2 (2001), 101–117. http://www.jstor.org/stable/2676784

[5] G. Casella and R. L. Berger. 2002. *Statistical Inference.* Thomson Learning. https://books.google.it/books?id=0x_vAAAAMAAJ

[6] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang. 2016. Data Cleaning: Overview and Emerging Challenges. In *Proc. of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016.* ACM, 2201–2206. https://doi.org/10.1145/2882903.2912574

[7] W. G. Cochran. 1977. *Sampling Techniques, 3rd Edition.* John Wiley. https://doi.org/10.1017/S0013091500025724

[8] T. Dalenius and J. L. Hodges. 1959. Minimum Variance Stratification. *J. Amer. Statist. Assoc.* 54, 285 (1959), 88–101. https://doi.org/10.1080/01621459.1959.10501501

[9] N. Dean and M. Pagano. 2015. Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys. *Journal of Survey Statistics and Methodology* 3, 4 (10 2015), 484–503. https://doi.org/10.1093/jssam/smv024

[10] O. Deshpande, D. S. Lamba, M. Tourn, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. 2013. Building, maintaining, and using knowledge bases: a report from the trenches. In *Proc. of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013.* ACM, 1209–1220. https://doi.org/10.1145/2463676.2465297

[11] D. Esteves, A. Rula, A. J. Reddy, and J. Lehmann. 2018. Toward Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis. *ACM J. Data Inf. Qual.* 9, 3 (2018), 16:1–16:26. https://doi.org/10.1145/3177873

[12] R. A. Fisher. 1954. *Statistical Methods for Research Workers, 12th Edition.* Edinburgh Oliver & Boyd.

[13] J. Gao, X. Li, Y. E. Xu, B. Sisman, X. L. Dong, and J. Yang. 2019. Efficient Knowledge Graph Accuracy Evaluation. *Proc. VLDB Endow.* 12, 11 (2019), 1679–1691. https://doi.org/10.14778/3342263.3342642

[14] J. Gao, Y. Xu, P. K. Agarwal, and J. Yang. 2021. Efficiently Answering Durability Prediction Queries. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021.* ACM, 591–604. https://doi.org/10.1145/3448016.3457305

[15] K. Gashteovski, S. Wanner, S. Hertling, S. Broscheit, and R. Gemulla. 2019. OPIEC: An Open Information Extraction Corpus. In *Proc. of the 1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019.* https://doi.org/10.24432/C53W2J

[16] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artif. Intell.* 194 (2013), 28–61. https://doi.org/10.1016/j.artint.2012.06.001

[17] R. V. Hogg, E. A. Tanis, and D. L. Zimmerman. 2013. *Probability and Statistical Inference.* Pearson. https://books.google.it/books?id=I_7tnQEACAAJ

[18] I. F. Ilyas, J. Lacerda, Y. Li, U. F. Minhas, A. Mousavi, J. Pound, T. Rekatsinas, and C. Sumanth. 2023. Growing and Serving Large Open-domain Knowledge Graphs. In *Companion of the 2023 International Conference on Management of Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023.* ACM, 253–259. https://doi.org/10.1145/3555041.3589672

[19] I. F. Ilyas, T. Rekatsinas, V. Konda, J. Pound, X. Qi, and M. A. Soliman. 2022. Saga: A Platform for Continuous Construction and Serving of Knowledge at Scale. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022.* ACM, 2259–2272. https://doi.org/10.1145/3514221.3526049

[20] G. Kalton. 1979. Ultimate Cluster Sampling. *Journal of the Royal Statistical Society. Series A (General)* 142, 2 (1979), 210–222. http://www.jstor.org/stable/2345081

[21] L. Kish. 1965. *Survey Sampling.* Wiley. https://books.google.it/books?id=xiZmAAAAIAAJ

[22] L. Kish. 1995. Methods for Design Effects. *J. Off. Stat.* 11, 1 (1995), 55. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/methods-for-design-effects.pdf

[23] E. L. Korn and B. I. Graubard. 1998. Confidence Intervals for Proportions with Small Expected Number of Positive Counts Estimated from Survey Data. *Survey Methodology* 24 (1998), 193–201. https://www150.statcan.gc.ca/n1/pub/12-001-x/1998002/article/4356-eng.pdf

[24] F. Li, X. L. Dong, A. Langen, and Y. Li. 2017. Knowledge Verification for Long-Tail Verticals. *Proc. VLDB Endow.* 10, 11 (2017), 1370–1381. https://doi.org/10.14778/3137628.3137646

[25] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. 2015. A Survey on Truth Discovery. *SIGKDD Explor.* 17, 2 (2015), 1–16. https://doi.org/10.1145/2897350.2897352

[26] N. G. Marchant and B. I. P. Rubinstein. 2017. In Search of an Entity Resolution OASIS: Optimal Asymptotic Sequential Importance Sampling. *Proc. VLDB Endow.* 10, 11 (2017), 1322–1333. https://doi.org/10.14778/3137628.3137642

[27] T. M. Mitchell, W. W. Cohen, E. R. Hruschka Jr., P. P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. D. Mishra, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. A. Platanios, A. Ritter, M. Samadi, B. Settles, R. C. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2018. Never-ending learning. *Commun. ACM* 61, 5 (2018), 103–115. https://doi.org/10.1145/3191513

[28] J. Mohoney, A. Pacaci, S. R. Chowdhury, A. Mousavi, I. F. Ilyas, U. F.q Minhas, J. Pound, and T. Rekatsinas. 2023. High-Throughput Vector Similarity Search in Knowledge Graphs. *Proc. ACM Manag. Data* 1, 2 (2023), 197:1–197:25. https://doi.org/10.1145/3589777

[29] R. G. Newcombe. 1998. Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine* 17, 8 (1998), 857–872. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

[30] P. Ojha and P. P. Talukdar. 2017. KGEval: Accuracy Estimation of Automatically Constructed Knowledge Graphs. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017.* ACL, 1741–1750. https://doi.org/10.18653/v1/d17-1183

[31] J. Pujara, E. Augustine, and L. Getoor. 2017. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017.* ACL, 1751–1756. https://doi.org/10.18653/v1/d17-1184

[32] Y. Qi, W. Zheng, L. Hong, and L. Zou. 2022. Evaluating Knowledge Graph Accuracy Powered by Optimized Human-Machine Collaboration. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022.* ACM, 1368–1378. https://doi.org/10.1145/3534678.3539233

[33] N. Queralt-Rosinach, J. Piñero González, À. Bravo, F. Sanz, and L. I. Furlong. 2016. DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. *Bioinformatics* 32, 14 (03 2016), 2236–2238. https://doi.org/10.1093/bioinformatics/btw214

[34] R. Reinanda, E. Meij, and M. de Rijke. 2020. Knowledge Graphs: An Information Retrieval Perspective. *Found. Trends Inf. Retr.* 14, 4 (2020), 289–444. https://doi.org/10.1561/1500000063

[35] M. Samadi, P. P. Talukdar, M. M. Veloso, and T. M. Mitchell. 2015. AskWorld: Budget-Sensitive Query Evaluation for Knowledge-on-Demand. In *Proc. of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015.* AAAI Press, 837–843. http://ijcai.org/Abstract/15/123

[36] S. E. Vollset. 1993. Confidence Intervals for a Binomial Proportion. *Statistics in Medicine* 12, 9 (1993), 809–824. https://doi.org/10.1002/sim.4780120902

[37] D. Vrandecic and M. Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. https://doi.org/10.1145/2629489

[38] S. Wallis. 2013. Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods. *J. Quant. Linguistics* 20, 3 (2013), 178–208. https://doi.org/10.1080/09296174.2013.799918

[39] J. Wang, S. Krishnan, M. J. Franklin, K. Goldberg, T. Kraska, and T. Milo. 2014. A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014.* ACM, 469–480. https://doi.org/10.1145/2588555.2610505

[40] X. Wang, L. Chen, T. Ban, M. Usman, Y. Guan, S. Liu, T. Wu, and H. Chen. 2021. Knowledge Graph Quality Control: A Survey. *Fundam. Res.* 1, 5 (2021), 607–626. https://doi.org/10.1016/j.fmre.2021.09.003

[41] Y. Wang, A. Khan, X. Xu, J. Jin, Q. Hong, and T. Fu. 2022. Aggregate Queries on Knowledge Graphs: Fast Approximation with Semantic-aware Sampling. In *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022.* IEEE, 2914–2927. https://doi.org/10.1109/ICDE53745.2022.00263

[42] E. B. Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212. https://doi.org/10.1080/01621459.1927.10502953

[43] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. 2016. Quality assessment for Linked Data: A Survey. *Semantic Web* 7, 1 (2016), 63–93. https://doi.org/10.3233/SW-150175