Uncertainty Discourse and Applications

www.uda.reapress.com

Uncert. Disc. Appl. Vol. 3, No. 3 (2026) 235–248.

Paper Type: Original Article

Fuzzy Data Quality Profiling with Decision Theoretic Treatment of Incomplete and Inconsistent Records

Bita Alipoor[1], Setayesh Zahiri[1], Seyyed Ahmad Edalatpanah[2,*], Duško Tešić[3]

[1] Morvarid Intelligent Industrial Systems Research Group, Iran.

[2] University College, Korea University, Seoul 02841, South Korea; s.a.edalatpanah@aihe.ac.ir.

[3] Military Academy, University of Defence in Belgrade, 11042 Belgrade, Serbia; tesic.dusko@yahoo.com.

Citation:

Received: 15 October 2025 *Revised: 20 January 2026* *Accepted: 23 March 2026*	Alipoor, B., Zahiri, B., Edalatpanah, S. A., & Tesic, D. (2026). Fuzzy data quality profiling with decision theoretic treatment of incomplete and inconsistent records. *Uncertainty Discourse and Applications, 3*(3), 235-248.

Abstract

Record-level data treatment requires a distinction between the degree of a quality defect and the loss attached to an intervention. We develop FQ-DTP, a fuzzy quality profiling and decision-theoretic framework that links completeness, consistency, and plausibility evidence to treatment actions. The general formulation separates graded memberships from latent-state probabilities and minimizes expected loss, with an optional query budget. Reported experiments evaluate a simpler five-action plug-in damage policy on an Adult subsample and three scikit-learn datasets with injected missingness and functional-dependency violations. At 30% Missing Not At Random (MNAR) missingness and 15% violations, mean cost-adjusted utility is 0.791, compared with 0.718 for mean/mode imputation, 0.710 for iterative-lite imputation, and 0.521 for a hard gate. Indicator-aware keep-all (0.882) and rule-based repair (0.844) remain stronger. Action agreement with an oracle using the same loss convention ranges from 83.6% to 97.4%, indicating internal agreement rather than external optimality. Ablations favour graded evidence and the specified asymmetric costs in this setting. The study does not evaluate the general posterior model, plausibility assessment, or query-budget formulation; predictive validation, clean-data calibration, query provenance, and clustered-seed inference require verification before deployment claims can be made.

Keywords: Fuzzy data quality, Data profiling, Decision theory, Missing data, Data cleaning, Bayes risk.

1 | Introduction

A missing field, a violated dependency, and a contradictory value impose different demands on a data treatment policy. Their practical consequences depend on the use of the record and on the cost of intervention. Data quality frameworks describe fitness for use [1], while measurement and improvement methods provide the evidence needed to diagnose defects [2]. The question addressed here is how to use graded evidence in an explicit record-level decision among retention, imputation, repair, source query, and discard.

Corresponding Author: s.a.edalatpanah@aihe.ac.ir

https://doi.org/10.48313/uda.v3i3.112

Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

Data profiling discovers statistical properties and dependencies [3]. Data cleaning supplies detection and repair mechanisms [4], [5]. Missing-data analysis distinguishes Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) and explains how their assumptions affect inference [6], [7]. These approaches already include decision and optimization methods; the present work contributes a particular connection between graded record evidence and a configurable treatment-loss model.

A binary violation flag can conceal differences in severity or attribute importance. Fuzzy membership functions provide a graded representation [8], while statistical decision theory specifies how uncertain states and asymmetric losses determine an action [9]. Combining them requires an inferential bridge: a membership degree is not itself a posterior probability. The general framework states this bridge explicitly and the experiment evaluates a restricted plug-in implementation.

We distinguish the objective of a repair operator from the objective of selecting an intervention. A procedure can reduce violations yet be costly or unnecessary for a given downstream use. The proposed framework makes action costs inspectable; its value is assessed against both active treatment policies and an indicator-aware policy that retains all records.

The framework defines graded completeness, consistency, and plausibility evidence, separates membership from state probability, and formulates expected-loss treatment with five named actions.

For additive record losses and a cardinality query budget, querying can be optimized by selecting at most B records with the largest positive net savings relative to their best immediate actions.

Algorithm 1 describes the general posterior-based framework. The experiments use a simpler completeness-and-consistency score and plug-in damage estimates, without evaluating its posterior or budget components.

The empirical contribution is the reported comparison and ablation on four public datasets. The strongest baselines are retained in the interpretation, and oracle agreement is confined to the shared loss convention.

The remainder of the manuscript is organized as follows. Section 2 reviews data quality frameworks, fuzzy uncertainty models, and decision-theoretic approaches to cleaning and imputation, and distills the research gap. Sections 3 and 4 formulate the problem and develop the proposed methodology, Section 5 reports the executed experimental study, Section 6 confronts limitations and counterarguments, and Section 7 concludes with a research roadmap.

2 | Related Work

2.1 | Data Quality Frameworks Dimensions and Profiling

Data-quality frameworks describe fitness for use across multiple dimensions [1]. Batini and Scannapieco consolidate dimensions and improvement methods [2]. These sources motivate application-dependent quality assessment; they do not supply numerical treatment costs for the experiment.

Profiling extracts metadata including cardinalities, functional dependencies, and unique column combinations [3]. Abedjan and colleagues evaluate error-detection tools on real datasets with ground-truth information, illustrating heterogeneous detection performance [5]. Cleaning research supplies operators and workflows [4]. These sources motivate evidence-based treatment; they do not validate the corruption generator used here.

2.2 | Fuzzy and Uncertainty Aware Models of Data Quality

Fuzzy membership assigns degrees in [0,1] to predicates such as complete or consistent [8]. Its grades can encode severity relative to rules without being probabilities of a latent state. That distinction is essential when the evidence is later consumed by expected-loss minimization.

Interval type-2 fuzzy sets represent uncertainty about membership degrees [10]. This is a possible extension when membership parameters are disputed, but the reported experiment uses point-valued completeness and

consistency scores. Label-noise research supplies context for uncertain supervised values [11], rather than a calibration procedure for the present quality scores.

Graded quality functions retain information about partial constraint satisfaction and evidential conflict. The present framework uses them as measurements consumed by a separate decision model. This design is one possible connection from representation to action, rather than a claim that fuzzy uncertainty research has not previously supported decisions.

2.3 | Decision Theoretic Cleaning Repair and Imputation

Expected-loss minimization is a standard decision principle [9], and cost-sensitive learning applies it to predictions [12]. Here the available actions operate on records before downstream modelling. Optimality is relative to the supplied state probabilities, losses, and feasible action set, and does not imply improved external predictive performance.

Constraint-based cleaning detects inconsistencies with dependencies [13] and treats violation context in repair [14]. HoloClean uses probabilistic inference for holistic repair [15], while NADEEF provides a platform for cleaning rules [16]. ActiveClean connects cleaning with statistical modelling [17]. These approaches already link evidence and intervention; the present framework adds a specified choice among five record actions under one loss model.

Missing-data mechanisms govern the assumptions needed for treatment [6], [7]. Chained imputation and flexible imputation provide practical approaches [18], [19]. The choice to impute a record is distinct from how its values are filled. The present comparison evaluates that treatment choice under specified costs, without asserting that all imputation methods optimize only fit or ignore downstream use.

2.4 | Synthesis and Research Gap

Table 1 summarizes the functions of the representative sources considered here. Profiling describes evidence, cleaning and imputation transform data, and decision theory formalizes action losses. FQ-DTP specifies an interface among these functions at record level. The comparison is conceptual and does not establish the absence of earlier cost-aware treatment policies.

Table 1. Positioning of the proposed framework relative to the main research streams surveyed in Section 2.

Research Stream	Quality Representation	Treatment Decision	Cost Model	Representative Works
Data quality frameworks and profiling	Mostly categorical	Organizational rules, fixed	Implicit	[1–3]
Fuzzy and uncertainty modeling	Graded memberships	Representation can inform decisions	None	[8], [10]
Constraint-based cleaning	Binary violations	Minimize violation count	Method-specific repair objective	[14], [15]
Missing-data imputation	Missing/present indicator	Impute by policy	Model-specific statistical assumptions	[6], [7], [19]
Cost-sensitive decision theory	Not record-level	Threshold on expected cost	Explicit, classification-focused	[9], [12]
Proposed framework	Fuzzy memberships per dimension	Expected-loss choice among five actions	Explicit, configurable loss	This work

3 | Problem Formulation and Research Questions

3.1 | Research Questions and Hypotheses

The research questions separate representation, treatment, and sensitivity. A treatment policy maps each record to one of five actions: keep, impute, repair, discard, or query-source.

RQ1 asks how schema rules, dependencies, and statistical profiles can be expressed as calibrated graded evidence for completeness, consistency, and plausibility.

RQ2 asks whether expected-loss treatment over {keep, impute, repair, discard, query-source} improves cost-adjusted utility relative to fixed policies under specified action losses.

RQ3 asks how treatment choices and utility respond to membership parameters, calibration error, and loss misspecification.

The following hypotheses guide evaluation without a verified preregistration record. The reported experiment addresses restricted versions of them, as described in Section 5.

H1 predicts that graded evidence improves treatment decisions relative to hard flags built from the same information. The reported utility ablation does not separately establish better corruption-detection accuracy.

H2 predicts that asymmetric action costs improve utility compared with a uniform-cost specification. It does not predict superiority over every fixed policy or identify a general relation to symmetric losses.

H3 predicts that moderate misspecification changes utility gradually. A parameter sweep and a defined misspecification scale are needed to test that statement; two discrete ablations are insufficient.

3.2 | System Model and Notation

Let T be a relational table with schema S = (A_1,…,A_m). Each record r has quality evidence e(r) and memberships μ_d(r) ∈ [0,1] for completeness, consistency, and plausibility. Latent states belong to $\mathcal{L}$ = {clean, incomplete, inconsistent, both}; the five actions are A = {keep, impute, repair, discard, query}. The loss function is denoted L(a,ℓ), avoiding confusion between the state set and loss. *Table 2* lists the notation.

Table 2. Principal notation.

Symbol	Meaning	Role in the Framework
T, S	Table with schema S = (A_1..., A_m)	Object of assessment
r	A record (row) of T	Unit of decision
D	Quality dimensions {completeness, consistency, plausibility}	Assessment axes
mu_d(r)	Membership degree of record r in dimension d	Graded quality evidence
e(r)	Evidence vector: violations, nulls, outlier scores	Observable input
$\mathcal{L}$	Latent states {clean, incomplete, inconsistent, both}	Hidden ground truth
P(l | e(r))	Posterior probability of state l given evidence	Inferential bridge
A	Actions {keep, impute, repair, discard, query}	Decision space
L(a, l)	Loss of action a under true state l	Economic interface
R(a | r)	Posterior expected loss (Bayes risk) of action a	Decision criterion
a*(r), B	Selected action; query budget	Decision output

The model separates evidence from state: violations, nulls, and outliers are observed, whereas whether a record is truly inconsistent is not. The posterior P(l | e(r)) is the inferential bridge between the two. Losses attach to action-state pairs rather than observations, which makes the framework sensitive to application economics.

3.3 | Formal Problem Statement

For the general formulation, sort the three dimension memberships in descending order, $b_1 \geq b_2 \geq b_3$, and choose weights $w_j \geq 0$ with $\Sigma_j w_j = 1$. The ordered weighted average is $\mu_q(r) = \Sigma_j w_j b_j$, as in *Eq. (1)* [20]. The weight attached to the lowest membership controls how strongly a poor dimension affects the composite. This score is evidence rather than a posterior or an action.

$$\mu_q(r) = \mathrm{OWA}_w(\mu_{comp}(r), \mu_{cons}(r), \mu_{plaus}(r)). \quad (1)$$

Treatment follows the Bayes-risk criterion of statistical decision theory [9]. For action a and record r, the posterior expected loss is defined by *Eq. (2)*, and the risk-minimizing action $a^*(r)$ is selected by *Eq. (3)*. At the dataset level, the objective minimizes cumulative expected loss subject to a query budget B, as stated in *Eq. (4)*. The four expressions chain evidence to a treated table.

$$R(a \mid r) = \Sigma_{\ell \in \mathcal{L}} L(a,\ell)\, P(\ell \mid e(r)). \quad (2)$$

$$a^*(r) = \arg\min_{a \in A} R(a \mid r). \quad (3)$$

$$\min_{\pi} \sum_{r \in T} R(\pi(r) \mid r) \text{ subject to } |\{ r : \pi(r) = \text{query} \}| \leq B. \quad (4)$$

Eq. (3) is optimal per record only when there is no coupling among actions. *Eq. (4)* adds a query-count constraint while retaining additive losses. Any fixed policy can be represented in this policy space, but it need not be an optimal or degenerate solution. The experimental policy uses plug-in damage estimates and a minimum aggregation rather than the general Ordered Weighted Averaging (OWA)/posterior construction.

4 | Proposed Methodology

4.1 | Fuzzy Quality Profiling from Rules and Statistical Evidence

The profiling stage begins from two sources mature deployments possess. The first is declarative knowledge: schema types, value ranges, referential rules, and denial constraints, including conditional functional dependencies discovered during profiling [3], [13]. The second is statistical: per-attribute distributions, quantiles, and co-occurrence structure estimated from the table itself. Each quality dimension converts a slice of this evidence into a membership degree, as itemized below.

Completeness membership decreases with the criticality-weighted fraction of missing applicable attributes.

Consistency membership decreases with detected violation severity and the importance assigned to the constraints.

Plausibility membership measures deviation from a statistical or domain profile. This dimension is proposed but not exercised in the experiment.

A trapezoidal membership can express full acceptance within a core and gradual decline to zero across a support. Core and support choices require domain evidence or training-data calibration. A range rule alone does not determine a plausible support outside the range. Interval memberships can represent uncertainty about these choices, but disagreement between evidence sources does not uniquely determine their endpoints or convert them into probabilities [10].

The OWA operator in *Eq. (1)* is monotone for nonnegative normalized weights [20]. Its weights control aggregation behaviour; monotonicity alone does not establish calibration or suitability for safety-critical use. The tested policy combines completeness and consistency by the minimum, a specific aggregation that must be distinguished from a general weighted average.

4.2 | Estimation of Latent State Posteriors

A posterior model could connect graded evidence to latent states using state-specific likelihoods and priors. A beta likelihood for memberships would require its parameters, boundary conventions at zero and one, and calibration data to be specified. Conditional independence of evidence components is a working modelling assumption, not a property of overlapping constraints. This posterior extension is not implemented or validated in the reported experiment.

Calibration should be evaluated on audited records separate from predictive evaluation. Including the missingness pattern can be informative under MAR, but does not by itself guarantee a correctly specified posterior; MNAR requires additional assumptions or sensitivity analysis [6], [7]. Membership values remain distinct from probabilities throughout. The experiment substitutes plug-in damage estimates for the proposed likelihood and prior machinery.

4.3 | Decision Theoretic Record Treatment

Keep retains the record under the downstream encoding. Impute fills missing attributes; repair corrects detected constraint violations. Discard removes a record, with an explicit loss for information removed. Query seeks an authoritative response and includes both acquisition cost and expected residual loss after the answer. Distinguishing imputation from repair makes the five-action formulation consistent with the experimental policy.

Losses price each action under each latent state. Their application meaning must be elicited rather than inferred from a quality grade. Query risk includes its fee and the expected loss of the best action after a response. The experiment fixes costs at one specified setting and uses uniform-cost and crisp-score ablations; it does not evaluate families of practitioner-elicited loss matrices.

Let $A_0 = A \setminus \{query\}$, $R_0(r) = \min_{a \in A_0} R(a|r)$, and let $R_q(r)$ include the query fee and expected post-answer loss. The net saving is $s(r) = R_0(r) - R_q(r)$. For independent actions, additive losses, a common unit query cost in the count constraint, and $|Q| \leq B$, *Eq. (4)* is solved exactly by selecting up to B records with the largest strictly positive $s(r)$. Every unqueried record takes its best action in A_0. This follows by exchanging any selected lower-saving record with an unselected higher-saving one. Unequal resource budgets or interacting query answers require a different optimization.

4.4 | Pipeline Architecture and Algorithm

Algorithm 1 describes the general posterior-based pipeline. Profiling supplies evidence, membership functions express its degree, calibration supplies state probabilities, and risk computation supplies actions. Positive net savings ration queries under the count budget. The tested plug-in policy does not execute all these stages.

Algorithm 1. General posterior based treatment with a query count budget.

Input: table T, constraints C, training profiles P, state likelihoods and priors, loss model, OWA weights w, query budget B, query-answer model
1) Obtain profiles and calibration parameters using information available for treatment
2) For each record r in T do
3) e(r) <- collect evidence from C and P
4) For each dimension d in {comp, cons, plaus} do
5) Mu_d(r) <- trapezoidal membership anchored on e(r)
6) End for
7) Mu_q(r) <- OWA_w(mu_comp(r), mu_cons(r), mu_plaus(r))
8) For each latent state $\ell \in \mathcal{L}$ compute its likelihood and prior product
9) Normalize across $\mathcal{L}$ to obtain $P(\ell|e(r))$
10) Verify that the likelihood normalization is defined
11) Compute $R(a|r)$ for immediate actions $a \in A_0$
12) Compute query risk $R_q(r)$, including fee and expected post-answer loss
13) $R_0(r) \leftarrow \min_{a \in A_0} R(a|r)$
14) Record the best immediate action

15) S(r) ← R_0(r) − R_q(r)
16) End for
17) Q ← up to B records with the largest strictly positive s(r)
18) Execute actions; append outcomes to the audit log
19) Update priors and noise parameters from query answers
Output: treated table T', audit log, per-record risk report

Per-record evidence extraction and membership evaluation cost O(m + |C_r|), posterior evaluation O(|$\mathcal{L}$|) for a fixed evidence model, and risk computation O(|A||$\mathcal{L}$|). Query selection costs O(n log n) by sorting or less with a bounded selection structure, and storing decisions or savings costs O(n) in addition to profiles and constraints. Profiling cost depends on the discovery algorithm and constraint language; general dependency discovery cannot be bounded solely by pairwise attribute checks [3]. Consequently, a universal constant-factor overhead claim is not established.

4.5 | Assumptions and Failure Conditions

The design leans on assumptions stated here so that the evaluation can probe them. They concern evidence, constraints, losses, and the query channel. Each assumption is paired with the failure condition expected if it does not hold.

Conditional independence is a modelling simplification; correlated constraints can produce miscalibrated posteriors.

Constraints must be defensible. Incorrect rules can make a valid record appear inconsistent.

Profiles must be relevant to the treated batch; drift can invalidate plausibility membership.

Losses must describe the intended application. A mathematically optimal action under misspecified loss can be practically harmful.

Query responses require a defined source, delay, reliability, and cost. An ideal answer simulated from known truth does not establish a usable operational query channel.

These assumptions define the general framework's scope. The reported ablations examine graded evidence and one cost perturbation; they do not test every assumption, posterior calibration, query reliability, or the general count-budget solution.

5 | Experimental Results and Analysis

5.1 | Setup

The manuscript reports four datasets: a 3,000-row complete-case Adult subsample, breast cancer (569 rows), wine (178 rows), and an 800-row stratified digits subsample [21], [22]. Clean realizations supply reference F1 scores, feature importance, and calibration data. F1 is the harmonic mean of precision and recall; macro F1 averages class-specific F1 scores. Numerical features are reported as standardized on the clean realization, and downstream logistic regression uses five-fold evaluation. The separation among the 15% calibration split, treatment fitting, and final evaluation folds is not specified sufficiently to establish an untouched holdout. Reported execution counts and timing require the missing experiment manifest rather than inference from the listed grid.

Each table carries one documented Functional Dependency (FD). Adult has the natural dependency education -> education-num, for which the repair rule is the majority mapping estimated on the corrupted table itself. The scikit-learn tables are extended with one derived attribute that is an exact deterministic function of primitive attributes, emulating the denormalized derived columns of production tables: breast_cancer and wine use D = (f_a + f_b)/2 over two named attributes, and digits uses the mean of two central pixels; the repair rule is exact recomputation from the primitives. Violations are injected before

missingness: on v percent of records (v in {5, 15}) the FD attribute is shifted by a random draw from 0.25 to 1.25 interquartile ranges, so the corrupted value is inconsistent with the dependency. Detection is rule-based without ground truth, and the observable violation severity is the deviation from the rule-implied value normalized by the attribute's interquartile range and capped at one.

Missingness is then injected at m in {10, 20, 30} percent under the three canonical mechanisms [6]. MCAR masks each cell by an independent Bernoulli draw. MAR gives every cell of a column a missing probability proportional to a logistic function of a fully observed carrier attribute (hours-per-week for Adult, the first feature elsewhere), calibrated to the marginal rate m; the carrier itself is never missing. MNAR is a deterministic lower-tail rule: per column, the cells with the lowest magnitude score are masked, where the score is the own value for numeric attributes and the own category frequency for categorical ones, so that under MNAR the smallest values and the rarest categories disappear. Every corrupted cell and record is logged, and the injection protocol is defined for this study.

The six baselines are indicator-aware keep-all, listwise deletion, mean/mode imputation, a three-round chained iterative-lite imputer, a hard gate, and rule-based FD repair. Iterative-lite uses logistic models for categorical targets and ridge models for numerical targets with five importance-ranked predictors; it is an implementation inspired by chained imputation rather than the mice package or a multiple-imputation inference analysis [18], [19]. FQ-DTP uses importance-weighted missing mass and graded FD severity, combines their quality scores by the minimum, and chooses among keep, impute, repair, query, and discard using plug-in damage estimates. Reported action costs are 0, 0.05, 0.10, 0.15, and 0.30, with loss scale $\lambda = 0.5$. Imputation and repair error rates use a 15% clean calibration split; feature importances use a clean reference model. These privileged clean calibration inputs bound the experiment's operational scope. Exact risk formulas and the information supplied by a query require the implementation record.

Cost-adjusted utility is $U = F1 - C$, with five-fold logistic-regression F1 and mean action cost over input records. Binary F1 is used for Adult and breast cancer, and macro F1 for wine and digits. This utility fixes a scale between predictive F1 and action fees; it is not a general monetary utility. Retention and action agreement with a post-hoc oracle provide additional outcomes. Evaluation on a treated or retained population can differ from evaluation on the original population, so the handling of discarded test records must be verified. Plausibility, posterior estimation, OWA aggregation, and budget-constrained querying are not evaluated.

Table 3. Datasets, functional dependencies, and clean-data reference F1 (mean over the ten per-seed realizations).

Dataset	Rows (per seed)	Features	FD (repair rule)	Target (F1 type)	Clean F1
Adult (UCI) [21]	3,000 (complete-case subsample)	6 numeric + 8 categorical (one-hot, ~100 columns)	Education -> education-num (majority mapping)	Income >50K (binary F1)	0.659
Breast cancer [22]	569	30 numeric + derived	D = (mean radius + mean texture)/2 (recompute)	Benign/malignant (binary F1)	0.983
Wine [22]	178	13 numeric + derived	D = (alcohol + flavanoids)/2 (recompute)	3 cultivars (macro F1)	0.979
Digits (subsample) [22]	800 (stratified)	64 numeric + derived	D = (pixel(3,3) + pixel(3,5))/2 (recompute)	10 classes (macro F1)	0.956

5.2|Main Comparison RQ1 RQ2

The primary condition is the harshest cell of the design, m = 30 percent MNAR with v = 15 percent FD violations. *Table 4* reports the cost-adjusted utility of all seven policies on all four datasets, and *Fig. 1* traces the utility of the seven policies across the missingness levels for Adult under MNAR. Aggregated over the

four datasets, FQ-DTP attains U = 0.791, which exceeds mean/mode imputation (0.718, +0.073), iterative-lite (0.710, +0.081), the hard gate (0.521, +0.270), and listwise deletion (-0.002), while trailing the passthrough keep-all baseline (0.882, -0.091) and the pure rule-repair policy (0.844, -0.053).

At the primary condition the reported mean F1 for FQ-DTP is 0.878. The utility comparison combines this predictive result with action costs and discarded records. Adult F1 is reported as 0.640 for FQ-DTP versus 0.536 for mean/mode and 0.543 for iterative-lite; digits F1 is reported as 0.936 with 88% retention. Listwise deletion has almost no retained records on the most incomplete tables. These outcomes should be interpreted conditional on the reported encoding and evaluation population, rather than attributing the policy ranking solely to better imputed values.

Table 4. Cost-adjusted utility U (mean +/- std over ten seeds) at the primary condition m = 30% MNAR, v = 15% FD violations.

Policy	Adult	Breast cancer	Wine	Digits
Keep-all (passthrough)	0.643 +/- 0.019	0.978 +/- 0.004	0.972 +/- 0.009	0.935 +/- 0.007
Drop-incomplete	0.377 +/- 0.111	0.200 +/- 0.259	-0.287 +/- 0.002	-0.300 +/- 0.000
Mean/mode imputation	0.487 +/- 0.032	0.890 +/- 0.008	0.755 +/- 0.018	0.739 +/- 0.014
Iterative-lite (MICE)	0.494 +/- 0.025	0.895 +/- 0.008	0.717 +/- 0.027	0.734 +/- 0.014
Hard quality gate	0.469 +/- 0.022	0.707 +/- 0.010	0.432 +/- 0.455	0.478 +/- 0.240
Rule-based FD repair	0.628 +/- 0.018	0.928 +/- 0.005	0.929 +/- 0.009	0.893 +/- 0.007
FQ-DTP (proposed)	0.559 +/- 0.021	0.897 +/- 0.003	0.876 +/- 0.021	0.832 +/- 0.007

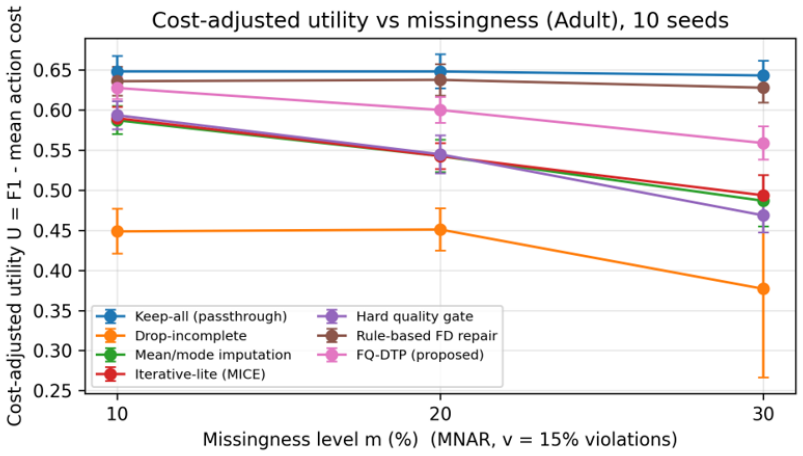


Fig. 1. Cost-adjusted utility versus missingness level m under MNAR with v = 15% FD violations (Adult; mean +/- std over ten seeds).

FQ-DTP dominates all imputing and filtering baselines at every level but is overtaken by the passthrough keep-all baseline and by pure rule-based repair.

5.3 | Ablation RQ2 RQ3

Table 5 compares the full policy with uniform action fees of 0.10 and with a crisp quality score. The uniform-cost variant retains an economic criterion but changes the relative price of keep, impute, repair, query, and discard. Both variants have lower displayed utility on all four datasets. On wine the uniform-cost deficit is 0.286; on digits the crisp-score deficit is 0.138. On Adult, 0.529 and 0.500 are the ablation outcomes, corresponding to deficits of 0.030 and 0.059 relative to 0.559. These separate perturbations support the usefulness of the stated design choices in this setting, but do not establish additivity, interaction effects, detection accuracy, or general robustness to misspecification.

Table 5. Ablation of FQ-DTP at the primary condition: Cost-adjusted utility (mean +/- std over ten seeds).

FQ-DTP Variant	Adult	Breast cancer	Wine	Digits
Full FQ-DTP	0.559 +/- 0.021	0.897 +/- 0.003	0.876 +/- 0.021	0.832 +/- 0.007
Uniform action costs (0.10 each)	0.529 +/- 0.020	0.890 +/- 0.002	0.590 +/- 0.452	0.773 +/- 0.025
Crisp (non-fuzzy) quality score	0.500 +/- 0.017	0.834 +/- 0.007	0.796 +/- 0.009	0.694 +/- 0.017

5.4 | Robustness across Missingness Mechanisms and Violation Levels RQ3

The reported mechanism scan gives mean utility under mild MCAR of approximately 0.832 for mean/mode, 0.836 for iterative-lite, and 0.853 for FQ-DTP. *Fig. 2* shows increasing intervention as missingness becomes more severe or informative: Adult query share rises from about 1.6% under 10% MCAR to about 33.5% under 30% MNAR. These patterns describe policy response under the injected mechanisms. They do not establish a general causal relation between missingness mechanism and treatment value, and the figure's action aggregates should be reconciled with *Table 8* before precise percentages are compared.

Table 6 compares 5% and 15% FD violations under 20% MCAR, a tripling of the violation rate. Displayed mean utility decreases by 0.012 for FQ-DTP, 0.003 for keep-all, and 0.005 for mean/mode. The limited change is consistent with the injection scheme: one derived or rule-governed attribute is perturbed and its repair is relatively direct. It does not demonstrate robustness to defects in primitive attributes, incorrect constraints, or larger violation rates.

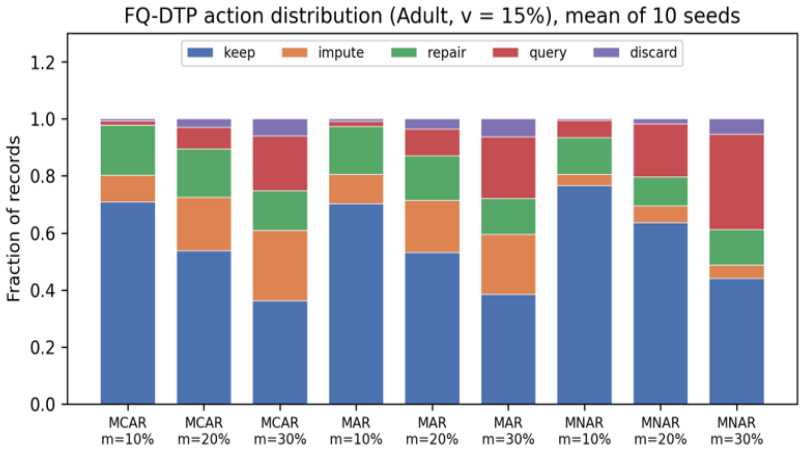


Fig. 2. Reported per-record action distribution on Adult across mechanisms and missingness levels at v = 15%, averaged over ten seeds.

Exact aggregation and query-answer provenance require verification; *Table 8* uses audit aggregates.

Table 6. Violation-rate axis at MCAR, m = 20%: Cost-adjusted utility and downstream F1, averaged over the four datasets (ten seeds).

Policy	U at v = 5%	U at v = 15%	F1 at v = 5%	F1 at v = 15%
Keep-all (passthrough)	0.857	0.854	0.857	0.854
Mean/mode imputation	0.808	0.803	0.857	0.852
Iterative-lite (MICE)	0.811	0.807	0.860	0.856
FQ-DTP (proposed)	0.825	0.813	0.862	0.859

5.5 | Statistical Analysis

Table 7 retains the reported Friedman and paired Wilcoxon statistics, with Holm correction for the six primary utility comparisons. Descriptive dataset means favour FQ-DTP over the two imputers, the hard gate, and listwise deletion, and favour keep-all and rule repair over FQ-DTP. The original tests treat four datasets times

ten seeds as 40 blocks. Seeds repeat related observations within datasets, so these p values do not establish cross-dataset significance as if 40 independent datasets had been sampled [23]. Raw outcomes, ties, validation splits, and the comparison family must be verified before inferential claims are made.

Table 7. Nonparametric tests over (dataset x seed) blocks: Friedman across the seven policies and paired Wilcoxon signed-rank tests [23–25], for FQ-DTP against each baseline on the primary metric.

Test	Statistic	p	Holm p	Outcome
Friedman, 7 policies, U at m=30% MNAR (40 blocks)	chi2(6) = 220.23	9.28e-45	-	significant
FQ-DTP vs keep-all (passthrough), U	W = 0.0	1.82e-12	1.09e-11	reject H0
FQ-DTP vs drop-incomplete, U	W = 1.0	3.64e-12	1.09e-11	reject H0
FQ-DTP vs mean/mode imputation, U	W = 6.0	2.55e-11	5.09e-11	reject H0
FQ-DTP vs iterative-lite (MICE), U	W = 19.0	5.58e-10	5.58e-10	reject H0
FQ-DTP vs hard quality gate, U	W = 0.0	1.82e-12	1.09e-11	reject H0
FQ-DTP vs rule-based FD repair, U	W = 0.0	1.82e-12	1.09e-11	reject H0
FQ-DTP vs rule-based FD repair, F1	W = 194.0	0.0031	-	reject H0
FQ-DTP vs keep-all (passthrough), U	W = 7.0	3.46e-11	-	reject H0

5.6|Audit Findings and Scope

Table 8 reports action agreement with a realized-loss oracle ranging from 0.836 on Adult to 0.974 on digits, and chosen-to-oracle mean-loss ratios from 1.006 to 1.070. The oracle uses known injected truth together with the policy's loss convention. These results measure internal decision agreement and regret under that convention. They do not certify optimal downstream F1, calibrated state probabilities, or the accuracy of an operational source-query service.

Indicator-aware keep-all has the highest reported utility on every dataset and rule-based repair is second. FQ-DTP pays for queries and other interventions under a damage convention that prices missing cells conservatively. Its lower utility is compatible with over-treatment relative to the downstream encoding, but the evidence does not isolate that mechanism from differences in calibration, retained populations, or encoding. The practical result is a benchmark requirement: a record-treatment policy should be compared with a strong missingness-aware retention baseline, not only with imputation and deletion.

The evidence covers four datasets, ten seeds, logistic regression, fixed action costs, and defects restricted to missingness and one FD attribute. The lower-tail MNAR mechanism is a particular self-masking design. General posterior calibration, plausibility assessment, practitioner loss elicitation, count-budget querying, and robustness to imperfect query answers remain untested. Clean-data feature importance and calibration are additional privileges. The reported results therefore establish descriptive comparisons within the synthetic corruption protocol rather than a generally optimal treatment policy.

Table 8. Audit at the primary condition: Action accuracy against the realized-loss oracle, mean realized-loss ratio, and chosen action distribution (aggregated over ten seeds and up to 200 audited records per seed).

Dataset	Action Accuracy vs Oracle	Realized-Loss Ratio	Keep/Impute/Repair/Query/ Discard (Chosen, %)
Adult	0.836 +/- 0.041	1.070	44.7 / 4.3 / 12.1 / 33.4 / 5.5
Breast cancer	0.924 +/- 0.022	1.033	47.0 / 0.0 / 8.2 / 43.7 / 1.0
Wine	0.918 +/- 0.019	1.025	43.4 / 1.8 / 3.9 / 48.8 / 2.1
Digits	0.974 +/- 0.014	1.006	42.9 / 0.7 / 0.5 / 45.2 / 10.7

The numerical results are reported at proof-of-concept scale. Independent regeneration requires code, raw outcomes, split indices, query records, and the execution manifest; a verified release is not specified.

6 | Discussion Limitations and Counterarguments

6.1 | Limitations

Elicitation is central to the framework: membership parameters, constraint severities, and action losses determine both evidence interpretation and treatment. The general calibration model offers a route for testing these choices, but it has not been exercised here. The two reported ablations show sensitivity to particular substitutions rather than a full robustness analysis.

Overlapping constraints can violate the independence assumption of the proposed posterior. Profile drift and incorrect rules can also misgrade records. Under MNAR, identifying a treatment model requires assumptions beyond the observed table [7]. These limitations motivate calibration and mechanism sensitivity studies; they are not resolved by assigning a fuzzy membership degree.

6.2 | Alternative Viewpoints and Responses

A probabilistically minded critic may ask why quality is modeled with membership functions at all, rather than with probabilities end to end. The answer is pragmatic: possibility-style grades tolerate partial constraint satisfaction and evidence conflict without a fully specified generative model, and the hybrid design already lets probabilistic posteriors consume whatever the grades encode. A data-management critic may counter that constraint-based cleaning already decides, by minimizing violations. It does, but violation count is a technique-internal objective; a repair that minimizes violations can still be the wrong economic choice for a record whose query cost is trivial.

A treatment policy learned from audit logs could provide a useful comparator where representative outcomes are available. Its auditability, sample requirements, and distribution shift should be evaluated rather than assumed inferior. FQ-DTP provides an explicit evidence-and-loss structure that can be examined or learned; its posterior component already requires priors, so it is not prior-free.

7 | Conclusion and Future Directions

FQ-DTP connects graded quality evidence to explicit record-treatment losses. The general framework distinguishes memberships from posterior probabilities and gives an exact query-count solution for additive independent risks. The reported plug-in experiment is narrower: completeness and consistency scores, five actions, clean calibration information, and fixed action costs. At 30% MNAR missingness and 15% FD violations, its mean utility is 0.791, exceeding mean/mode (0.718), iterative-lite (0.710), the hard gate (0.521), and listwise deletion, while trailing indicator-aware keep-all (0.882) and rule repair (0.844). The utility ablations favour graded evidence and the specified asymmetric costs. Oracle agreement of 83.6–97.4% establishes agreement under a shared loss convention, not external optimality.

The strongest retention and repair baselines make calibration against downstream use the next priority. A confirmatory study should separate clean calibration data from evaluation, compare policies on a common untouched population, specify all five risk estimates, and document query answers and their costs. It should then test plausibility, the posterior model, budget constraints, and realistic primitive-attribute corruption. These steps would determine whether the explicit treatment structure improves decisions beyond the controlled setting reported here.

Author Contributions

B. A.: Conceptualization, methodology, software, and writing original draft. S. Z.: Data curation, formal analysis, and visualization. A. A. S.: validation, supervision, and writing review and editing.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

Adult is available from the UCI Machine Learning Repository; breast cancer, wine, and digits are distributed with scikit-learn. Exact subsamples, corruption realizations, derived attributes, and splits require the experiment records.

Code Availability

A verified public release of the experiment code and raw outcomes is not specified.

Generative AI Disclosure

AI-assisted tools were used in drafting and revising the manuscript text. The authors retain responsibility for verification of the scientific content and the applicable disclosure.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

[2] Batini, C., & Scannapieco, M. (2016). *Data and information quality: Dimensions, principles and techniques* (Vol. 13). Springer. https://doi.org/10.1007/978-3-319-24106-7%0A%0A

[3] Abedjan, Z., Golab, L., & Naumann, F. (2015). Profiling relational data: A survey. *The VLDB journal*, *24*(4), 557–581. https://doi.org/10.1007/s00778-015-0389-y%0A%0A

[4] Ilyas, I. F., & Chu, X. (2019). *Data cleaning*. Morgan & Claypool. https://doi.org/10.1145/3310205

[5] Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., & Tang, N. (2016). Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB endowment*, *9*(12), 993–1004. https://doi.org/10.14778/2994509.2994518

[6] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. http://dx.doi.org/10.1093/biomet/63.3.581

[7] Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons. https://doi.org/10.1002/9781119482260%0A%0A

[8] Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

[9] Berger, J. O. (1987). Statistical decision theory. In *The new palgrave dictionary of economics* (pp. 1–6). Springer. https://doi.org/10.1057/978-1-349-95121-5_1872-1%0A%0A

[10] Mendel, J. M., & John, R. I. B. (2002). Type-2 fuzzy sets made simple. *IEEE transactions on fuzzy systems*, *10*(2), 117–127. https://doi.org/10.1109/91.995115

[11] Frénay, B., & Verleysen, M. (2013). Classification in the presence of label noise: A survey. *IEEE Transactions on neural networks and learning systems*, *25*(5), 845–869. https://doi.org/10.1109/TNNLS.2013.2292894

[12] Elkan, C. (2001). The foundations of cost-sensitive learning. *International joint conference on artificial intelligence* (Vol. 17, pp. 973–978). Lawrence Erlbaum Associates Ltd. https://cseweb.ucsd.edu//~elkan/rescale.pdf

[13] Bohannon, P., Fan, W., Geerts, F., Jia, X., & Kementsietsidis, A. (2007). Conditional functional dependencies for data cleaning. *2007 IEEE 23rd international conference on data engineering* (pp. 746–755). IEEE. https://doi.org/10.1109/ICDE.2007.367920

[14] Chu, X., Ilyas, I. F., & Papotti, P. (2013). Holistic data cleaning: Putting violations into context. *2013 IEEE 29th international conference on data engineering (ICDE)* (pp. 458–469). IEEE. https://doi.org/10.1109/ICDE.2013.6544847

[15] Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2017). Holoclean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment* (pp. 1190-1201). ACM. https://doi.org/10.14778/3137628.3137631

[16] Dallachiesa, M., Ebaid, A., Eldawy, A., Elmagarmid, A., Ilyas, I. F., Ouzzani, M., & Tang, N. (2013). NADEEF: A commodity data cleaning system. *Proceedings of the 2013 ACM sigmod international conference on management of data* (pp. 541–552). ACM. https://doi.org/10.1145/2463676.2465327

[17] Krishnan, S., Wang, J., Wu, E., Franklin, M. J., & Goldberg, K. (2016). Activeclean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB endowment* (pp. 948–959). ACM. https://doi.org/10.14778/2994509.2994514

[18] Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*, 1-67. https://doi.org/10.18637/jss.v045.i03

[19] van Buuren, S. (2018). Flexible Imputation of Missing Data, Second Edition (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259

[20] Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE transactions on systems, man, and cybernetics*, *18*(1), 183–190. https://doi.org/10.1109/21.87068

[21] Kelly, M., Longjohn, R., & Nottingham, K. (2023). *The UCI machine learning repository*. https://archive.ics.uci.edu/

[22] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*, 2825–2830. https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[23] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of machine learning research*, *7*(Jan), 1–30. https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

[24] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

[25] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70. https://doi.org/10.2307/4615733