PEER REVIEW RECORD

FlexRL: an R package for Flexible Record Linkage and Linked Data Quality Assessment

Kayané Robach, Mark A. van de Wiel and Michel H. Hof

Vol. 1 (2026), Art. 11 · https://doi.org/10.68365/jrepstat.2026.11

The decision letters, the reports the authors read, their answers and the editor's signed decisions, as written. The reports come from the journal's review engine. The authors could have kept this record private until they confirmed the proof.

DECISION PATH

2026-10-23	Major revision
2026-10-26	28 answers
2026-10-26	Minor revision
2026-10-26	8 answers
2026-10-28	Accepted

Round 1: Major revision

SIGNED DECISION

Confirmed and signed by Matthias Templ on 2026-10-23 · ORCID 0000-0002-8638-5276

DECISION LETTER

Thank you for your submission to JRepStat. We are inviting a major revision. The reviewers' points follow, grouped by weight.

FlexRL packages a hierarchical two-source record linkage model (StEM) together with method-agnostic diagnostics, two FDP estimators and several sample-divergence measures, so a practitioner linking sources without a unique identifier can judge how the linkage decision rule affects downstream inference. It matters to statistical agencies and to researchers working with administrative, registry and survey data. We would be glad to see the revised version.

How to read this letter. The decision rests on the items marked required. If a required item seems wrong to you, or cannot be done within this paper, decline it on the revision page and give the reason. A reasoned decline is a legitimate answer, and the editor weighs it. Reviewers can be mistaken. Items marked recommended or optional are advice from the reviewers. Adopt them, decline them with a sentence, or leave them for later work. None of that counts against the paper. The reviewer reports below carry the same marks. The editor who signs this decision may weigh an item differently in the accompanying message, and that message prevails.

REQUIRED (the decision rests on these):

- [consistency-check] Table 7 does not reconcile internally. Taking BRL, sensitivity from the table is TP/(TP+FN) = 80/(80+438) = 0.15, but the table reports 0.26; its FDP from FP/(TP+FP) = 21.90/(80+21.90) = 0.21, reported as 0.23. The same mismatch holds for multilink (sensitivity 137/517 = 0.27 vs 0.36; FDP 270.70/407.70 = 0.66 vs 0.58), fastLink, fedmatch and FlexRL. Please reconcile the reported rates with the counts, or state the aggregation rule in the caption. Tables 2, 3 and 6 reproduce their rates exactly from their counts, so Table 7 is the outlier.
- [consistency-check] Table 6 and Table 7 appear to be on different scales. Table 6 (no blocking) implies 10568 true links (TP+FN, consistent across fastLink, fedmatch, FlexRL). Table 7 (with blocking) shows TP+FN of about 517 to 518 per method, roughly one twentieth of 10568, which matches a per-region average over the 20 regional blocks. Section 5 states region is stable and error free with agreement 1.00, so blocking on region should not drop true links from the national total. Please state

whether Table 7 reports per-region averages or totals, since the two tables are presented as a direct comparison. A one-sentence caption note resolves it.

- [simulation-design] No replications or Monte Carlo error are reported for the headline simulation tables. Table 2 and Table 4 appear to come from a single simulated pair of files (350/400 records). Several reported gaps are within plausible sampling noise of one run, for example FlexRL dynamics sensitivity 0.59 against fedmatch 0.60, and FlexRL FDP 0.16 against multilink 0.16. The FDP estimators are averaged over n_repeats, but the TP/FP/sensitivity/FDP comparison is not. Repeat the generate-and-link cycle across replications and report the spread of sensitivity and FDP, or qualify the dynamics claim to a single illustrative run and drop any ranking of methods that are within one run of each other.
- [ethics-and-integrity] Add a competing-interests statement. The authors are the developers of FlexRL (Robach and Hof 2026) and the paper benchmarks that package against competing methods in Tables 2, 3, 6 and 7, so a disclosure of this role is a standard integrity obligation that is currently absent.
- [Reviewer 1] In the SHIW application the score-based FDP estimate at the chosen threshold is about 0.03 while the true FDP is 0.16. State this gap explicitly where the print output appears, and note the synthetic estimate (0.12) was closer, so a practitioner is warned which estimator held up here.
- [Reviewer 3] Document how random draws are produced inside the OpenMP-parallelised Rcpp samplers and whether set.seed() reproduces a run bit-for-bit when threads > 1. If it does not, qualify the reproducibility claim and state the thread count used for the reported tables.
- [Reviewer 3] For each benchmark table (2, 3, 6, 7), record the seed(s), R version, OS and hardware, and the version of every compared package, or state these in the text. The comparative timing, memory and performance numbers depend on all of them.

RECOMMENDED:

- [simulation-design] The simulation DGP is the package's own simulate_data() function, which synthesises records from the same hierarchical generative model FlexRL fits (registration errors, missingness, a dynamic PIV governed by a hazard). This is a DGP aligned with the method under test, and the comparison-vector competitors are evaluated on data generated under assumptions they do not share. The headline claim that "modelling PIVs dynamics improves record linkage" rests on exactly this self-generated setting (Table 2: FlexRL dynamics sensitivity 0.59 vs no-dynamics 0.49). Report the same comparison on data generated outside the FlexRL model, or state in Section 2.4 that the simulation is a demonstration of the model's own DGP and that the dynamics benefit is quantified on real linkage only.
- [simulation-design] The simulation factor grid is a single configuration: one sample size, one overlap (300 of 350/400), 2% mistakes, 0.5% missingness, four PIVs, one dynamic PIV. The conditions that would stress the method and expose its failure modes, lower overlap, higher error rates, fewer discriminative PIVs, more dynamic PIVs, are not varied in the simulation. The real-data sections add scale and a national/regional contrast, but not a controlled sweep of these factors. Vary overlap and error rate across a small grid and report how the dynamics advantage and the FDP estimates hold up.
- [methods-soundness] The threshold ξ=0.82 is chosen in Section 4.4 from the divergence "elbow", and in Section 4.1/Figure 5 that choice is then shown to give a regression estimate close to the known truth. In the SHIW the truth is available, so this is a validation of the elbow heuristic, but the heuristic itself is the only guidance a practitioner has when the truth is unknown. State explicitly that the elbow rule is a heuristic and that its reliability away from benchmark data where true links exist has not been characterised.
- [methods-soundness] The cross-method tables (Table 2, Table 3, Table 6, Table 7) compare methods at the single default rule ξ=0.5. Methods with different score scales and operating points are not necessarily comparable at one fixed cutoff, as the wide spread in sensitivity and FDP shows (for example reclin2 at sensitivity 0.80 with FDP 1.00 versus FlexRL at 0.33 and 0.26 in Table 3). Note in the table captions or surrounding text that these are single-threshold snapshots and that the FDP-versus-threshold curves give the fuller comparison.
- [methods-soundness] The Discussion describes FlexRL as a "reliable solution to record linkage". On the SHIW the method reaches sensitivity near 0.33 to 0.39 with FDP near 0.26 to 0.57, which reflects the difficulty of the task rather than a fault, but the wording overstates what the numbers show.

Soften to describe the demonstrated behaviour, for example that the method controls FDP at a stated threshold while recovering a known fraction of links.

- [novelty-and-prior-art] The second stated contribution is a method-agnostic toolkit of linked-data quality diagnostics. Edwards et al. 2026 ("Assessing Linkage Quality: Metrics, Insights, and a Reusable Python Toolkit", IJPDS) addresses the same problem of quality metrics for linked data in a reusable toolkit and is not cited. This is a live, actively-published line (the manuscript already cites Laidler 2025, Lam 2024, Harron 2017/2020 in the same venue). Please position the diagnostics contribution against it. The published metrics are the point here; comparison of the software itself sits with another assessor.
- [software-quality] The DESCRIPTION declares BRL, arf, diyar, fastLink, fedmatch, mice, multilink, reclin2 and synthpop as hard `Imports`. These are needed only by the optional `link_with_<pkg>()` wrappers and the synthetic-data FDP estimator, so loading FlexRL for its own StEM linkage pulls in nine extra packages, and a single install failure among them blocks FlexRL. Moving the wrapper and synthesis dependencies to `Suggests` with `requireNamespace()` guards at the call sites would lighten the install and match how wrappers are used.
- [software-quality] The log-likelihood sanity check in `simulateD()` compares `round(Dsample$loglik, 3) != round(ll_check, 3)` and calls `stop()` on mismatch. Equality after rounding to three decimals can false-trigger on large-magnitude log-likelihoods where a benign floating-point difference exceeds 0.0005, and can mask a real discrepancy below that scale. A relative tolerance (for example `isTRUE(all.equal(..., tolerance = 1e-6))`) would be more robust.
- [statistical-rigor] The FDP estimates in the print output and tables are given as point values (for example 0.03 at threshold 0.82 against a true FDP of 0.16). Since the synthetic-data estimator is already averaged over n_repeats draws, reporting the dispersion across those repeats would show how much of the gap between estimated and true FDP is sampling variability of the estimator.
- [visualisation] Figure 5 distinguishes four regression fits by colour alone: the text names "estimate in dashed orange", "estimate in yellow" and "truth represented in red". Orange and red read as close hues under the common red-green deficiencies, and yellow lines carry low contrast on a white ground. Redraw Figure 5 with a colourblind-safe palette and pair each line with a distinct line style or a direct label, so the three thresholds (0.55, 0.82, 0.99) and the truth can be told apart without relying on colour.
- [writing-quality] One sentence in Section 3.5 reads "so the sweep is linear the number of candidate pairs", which is missing a word ("linear in the number"). The table column heading "FlexRL (no dynamic)" in Tables 2 and 4 would read better as "(no dynamics)" to match the "dynamics" wording used elsewhere.
- [ethics-and-integrity] State the provenance and access terms of the SHIW data in one sentence. The 2012 and 2014 waves are anonymised microdata released by the Bank of Italy for research, which places this work within secondary analysis of public data. Recording that status makes clear why separate ethics approval and participant consent are not required here.
- [Reviewer 1] Qualify the word "reliable" in the Discussion given the reported FDPs (0.57 on the full data, 0.29 with blocking); say that the scalability claim concerns compute and memory, not linkage accuracy.
- [Reviewer 1] Report the regression specifics: which SHIW variable encodes expected age of retirement, the linked sample size used (273 at the chosen threshold), the sensitivity there (0.26), and that the reference line is the estimate computed on true links.
- [Reviewer 1] State how the PIV independence violation (Occupation and Sex associated, Cramér's V 0.32 and 0.35) may affect calibration of the FDP estimators, beyond issuing a warning.
- [Reviewer 1] Give practitioners guidance on which FDP estimator to trust when the two disagree, since the SHIW example shows them diverging from each other and from the truth.
- [Reviewer 2] Clarify the scaling factor N_B/N_synth in Equation 11 and state when the synthetic estimator is expected to be biased. The simulation returns 0.28 against a true 0.16 for the dynamics model, which the text does not address.
- [Reviewer 2] Note that the linked-sample divergence measures are heuristic indicators of sampling bias, not estimates of the bias in β, and that the "elbow" threshold choice carries no formal guarantee.

- [Reviewer 3] Provide a single master script that maps each table and figure to the command that regenerates it, with a captured session-info / environment record.

OPTIONAL:

- [software-quality] The `.onAttach` startup message ("please cite us … feel free to complain as well") and the `music_on` argument that opens a browser tune are informal side effects. They are harmless and default off, but a plainer startup message reads better in a statistical-agency setting and avoids friction with package-check gates that flag chatty attach messages.
- [Reviewer 2] A broader simulation varying overlap, PIV discriminative power and error rates would map out when each FDP estimator is reliable.

We will correct these in production, and you confirm them in the proof:

- [consistency-check] The reference "Robach, Kayané, Mark A van de Wiel, Michel H Hof, and Stéphanie L van der Pas. 2026. Causal Record Linkage" ends with a truncated URL ("h."), and the Lam et al. 2024 entry carries a concatenated DOI ("…238910.23889/ijpds.v7i3.1794"). Confirm the correct links at proof.

REVIEWS

Report Reviewer 1, recommends minor

Written by the journal's review engine in the role of the application reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

The problem this package addresses is real and the framing is right for practice. Agencies and researchers routinely link survey, registry and administrative sources without a unique identifier, and the downstream analyst then works with a sample that carries two unobserved errors at once. A tool that returns FDP estimates and sample-divergence measures as functions of the decision rule, and that works on the output of any linkage method, is genuinely useful. The SHIW example is a good choice because the true links can be reconstructed, so the diagnostics can be shown against a known answer. The package is positioned clearly against the existing R landscape in Table 1, and the distinction between the comparison-vector paradigm and the hierarchical model is drawn fairly.

My main application concern is the honesty of the headline diagnostic in the worked example. At the chosen threshold of 0.82 the print output reports a score-based FDP of about 0.03, while the true FDP at that threshold is 0.16. A practitioner reading the score estimate alone would conclude the linked sample is almost clean when roughly one in six linked pairs is wrong. The paper already states that the score estimator is only as good as the linkage model and that calibration suffers under missing values, low overlap and weak PIVs, so the groundwork is there. What is missing is the explicit connection at the point where the numbers appear. The synthetic estimate (0.12) came closer to 0.16, which is the more reassuring result for your method, and saying so would help rather than hurt the package. Without that sentence the example reads as if the tool worked when its primary estimator underestimated the error it exists to flag.

A second application point concerns what the chosen threshold costs. At 0.82 the sensitivity is 0.26, so the regression runs on a heavily selected quarter of the recoverable links. This is exactly the sampling-bias direction the Section 4.3 diagnostics are meant to surface, and the elbow heuristic lands the user there by design. The reader would benefit from seeing the linked sample size (273), the sensitivity, and a line on how representative that remaining sample is, stated alongside the regression figure. The retirement-age regression itself needs a little more detail: name the SHIW variable, and state that the reference line is the estimate on true links, so the comparison in Figure 5 is unambiguous.

On the broader claims, "reliable solution" in the Discussion sits awkwardly next to the FDPs of 0.57 on the full data and 0.29 with blocking. The scalability demonstration is convincing as a statement about memory and running time on a standard machine, and that is a real strength. Separating the compute claim from an accuracy claim would make the Discussion match the tables. Finally, the independence assumption is violated in the very example used throughout (Occupation and Sex are associated, and working status and branch were merged for the same reason). The warning mechanism is good, but a sentence on what that dependence does to the calibration of the FDP estimators would set expectations for users who will

meet the same situation. These are matters of qualifying claims and connecting numbers already in the paper, not of new work.

Numbered suggestions to the authors:

1. REQUIRED Where the print output at threshold 0.82 is shown, state that the score-based FDP estimate (about 0.03) is below the true FDP (0.16) and that the synthetic estimate (0.12) was closer, so readers know which estimator held up in this application.
2. [editor_fix] Complete the reference "Robach et al. 2026. Causal Record Linkage. h.".
3. RECOMMENDED In the Discussion, qualify "reliable" and separate the compute-and-memory claim from linkage accuracy, given the reported FDPs of 0.57 (full data) and 0.29 (blocked).
4. RECOMMENDED Report the regression specifics: the SHIW variable for expected age of retirement, the linked sample size (273) and sensitivity (0.26) at the chosen threshold, and that the reference line is computed on true links; add one line on the representativeness of the retained sample.
5. RECOMMENDED State how the PIV independence violation (Occupation and Sex, Cramér's V 0.32 and 0.35) may affect the calibration of the FDP estimators, beyond the warning.
6. RECOMMENDED Add guidance on which FDP estimator a user should trust when the two disagree, since the SHIW example shows divergence between them and from the truth.

Report Reviewer 2, recommends minor

Written by the journal's review engine in the role of the methodology reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

This is a software paper built on methodology that is already published (the linkage model in Robach, van der Pas et al. 2025, the FDP estimators in Robach, Hof et al. 2025). I read it as a methodology reviewer, so my question is whether the statistics shown here are correct and presented so that a practitioner uses the tools as intended. The modelling is sound and the package fills a real gap: Paradigm 2 (hierarchical generative modelling) for linkage between two sources under a one-to-one constraint, with time-varying PIVs, was not previously available in R. The diagnostic layer that turns a linkage threshold into FDP and sampling-bias summaries is a genuine contribution for applied work. The estimation description in Section 3.2 is clear, the factorisation in Equation 3 and the StEM scheme in Algorithm 1 are coherent, and the inference decomposition in Equations 6 to 9 gives a usable account of the two biases.

My main concern is calibration of the headline tool in the paper's own illustration. In Figure 4 the true FDP at the selected threshold ξ=0.82 is 0.16, the sensitivity is 0.26, and the score-based estimate reads 0.03. The score-based estimator therefore understates the true FDP by about a factor of five in the example used to sell it. The text says in passing that this estimator is "only as good as the linkage model", but it does not connect that caveat to the number on the page. A reader who follows the recommended workflow (pick the elbow, check the estimated FDP is "reasonable") would see 0.03 and conclude the linked data are nearly clean when roughly one linked pair in six is spurious. Please state the discrepancy where it occurs and explain it: posterior scores that are overconfident drive the estimate down, and this is the motivation for the synthetic estimator, which gives 0.12 at the same threshold and sits closer to the truth. This is a short textual fix and the package is better for being honest about where its cheaper estimator fails.

The estimators are reported as point values with no spread. The procedure averages over n_repeats synthetic draws and over posterior draws, so the ingredients for an interval are already computed. For a tool whose purpose is to pick a threshold from an estimated FDP curve, the width of that estimate is part of the answer, and I would like to see it reported. Relatedly, the scaling in Equation 11 (multiplying the synthetic false-positive count by N_B/N_synth) assumes synthetic records are falsely linked at the same rate as real non-matching records. The simulation in Table 4 shows the synthetic estimate at 0.28 against a true 0.16 for the dynamics model, while the no-dynamics model gives 0.17 against 0.16. That pattern suggests the synthetic generator, which does not reproduce temporal change, mischaracterises the false-positive distribution when dynamics are modelled. A sentence stating the scaling assumption and the conditions under which the estimator is biased would let users read Table 4 correctly.

Two smaller methodological points. The divergence measures (SMD, support IoU, MMD, agreement rates) are descriptive indicators of how the linked sample differs from its source. They are not estimates of the bias in β, and the "elbow" rule for choosing the threshold is a visual heuristic with no formal guarantee. Both are reasonable for a practical tool, and both should be labelled as such so the user does not

over-read a flat MMD curve. Finally, in the SHIW analysis Sex and Occupation stay in the model despite the association the package itself flags (Cramér's V around 0.32 to 0.35), which violates the conditional-independence assumption that the model and the score-based FDP both rest on. A note on the likely direction of its effect would close the loop between the warning the software prints and the analysis the paper runs.

I could not see the vignettes or the reproducibility artifacts referenced in the text, so I have raised no ask for content that may sit inside them, and the reference issues I note (the truncated "h." URL, the concatenated DOI, the JASP version mismatch) are for the production list.

Numbered suggestions to the authors:

1. REQUIRED Where Figure 4 appears, state that the score-based FDP estimate at ξ=0.82 is 0.03 while the true FDP is 0.16, and qualify the calibration claim: the score-based estimator understates FDP when the posterior scores are overconfident, which is the reason the synthetic estimator (0.12 here) is provided.
2. [editor_fix] Complete the truncated reference "Robach et al. 2026, Causal Record Linkage" (URL reads "h."), fix the concatenated DOI in Lam et al. 2024, and reconcile the JASP version (text says 1.0, the reference says 0.98.8).
3. RECOMMENDED Report the spread of the FDP estimates and the divergence measures across the n_repeats synthetic draws and the posterior draws, so the threshold-dependent FDP comes with a measure of precision.
4. RECOMMENDED Clarify the N_B/N_synth scaling in Equation 11 and state when the synthetic estimator is expected to be biased, referencing the simulation where it returned 0.28 against a true 0.16 for the dynamics model.
5. RECOMMENDED Label the linked-sample divergence measures as heuristic indicators of sampling bias, not estimates of the bias in β, and note that the elbow threshold choice has no formal guarantee.
6. RECOMMENDED In the SHIW analysis, note that Sex and Occupation remain associated (Cramér's V around 0.32 to 0.35) against the conditional-independence assumption, and discuss the likely effect on linkage and on the FDP estimates.
7. OPTIONAL A broader simulation varying overlap, PIV discriminative power and error rates would characterise when each FDP estimator is reliable.

Report Reviewer 3, recommends minor

Written by the journal's review engine in the role of the software reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

This is a software paper, so I read it for whether the package can be run, whether the implementation is sound, and whether the reported numbers are the code's numbers. The engineering account is clear and the design choices are explained at the level a user needs. The workflow (prepare_data, naive_linkage, StEM, RL_diagnostics with print and plot S3 methods) is coherent, the Rcpp backend and the memory argument (linear in NA+NB and in candidate pairs, not in NA×NB) are stated precisely, and the function arguments are documented in the prose. Availability on CRAN, a GitHub repository with vignettes, and a JASP module is a strong distribution story. The diagnostics being agnostic to the linkage method, with wrappers for six other packages, is the kind of feature that makes a package useful beyond its own model.

My main concern is reproducibility of the stochastic core under parallelism. The paper states that all random draws use R's generator so that set.seed() makes a run reproducible, and separately that the samplers are parallelised over OpenMP threads with a default of four. R's RNG is not thread-safe, and floating-point summation order across threads is not fixed, so a reader cannot tell from the text whether a run with threads > 1 is bit-reproducible, reproducible up to Monte Carlo noise, or reproducible only when forced to a single thread. This matters for a package whose output is a stochastic linkage matrix and whose headline tables report exact counts. Please state how the draws are handled across threads and what set.seed() actually guarantees, and give the thread count used for the paper's tables.

The second concern is the provenance of the benchmark tables. Tables 2, 3, 6 and 7 compare FlexRL against six other packages on sensitivity, FDP, time and memory, and the scalability claim rests on them. Those numbers depend on the seed, the machine, the R version and the exact version of each compared package, and comparative timing in particular can shift substantially across package versions. The paper

gives citation years but not installed versions, and gives no seed or hardware specification for the reported runs. Please record these, in the text or in a reproduction script, so a reader can regenerate the reported figures or at least understand why their numbers differ. I note the memory caveat (R garbage collector, external allocations underestimated) is already disclosed, which is the right thing to do.

One reference, "Robach et al. 2026. Causal Record Linkage. h.", has a URL that extraction renders as a single "h." and is clearly incomplete. I have flagged it for the production list rather than as author work.

I could not run anything, because I was handed the manuscript and intake facts only. If the vignettes already carry a reproduction path for the comparative tables, some of the above is satisfied and the ask reduces to pointing at it and recording the environment. I have written the asks so they collapse to that if the material exists.

Numbered suggestions to the authors:

1. REQUIRED State how random draws are generated inside the OpenMP-parallelised Rcpp samplers and whether set.seed() reproduces a run bit-for-bit with threads > 1. If reproducibility holds only at a single thread or only up to Monte Carlo noise, say so in Section 3.5 and give the thread count used for the reported tables.
2. REQUIRED For Tables 2, 3, 6 and 7, record the seed(s), R version, operating system and hardware, and the installed version of every compared package, either in the text or in a reproduction script, so the comparative performance, timing and memory numbers can be regenerated or placed in context.
3. [editor_fix] Repair the truncated reference "Robach et al. 2026. Causal Record Linkage. h." with its full citation and URL.
4. RECOMMENDED Add one master script that maps each table and figure to the command producing it, together with a captured session-info or environment record, so a reader can reproduce the article's results end to end.

REVISION, RECEIVED 2026-10-26

Authors confirm each required point as done or give their reason for declining it. Further answers are voluntary, because the review compares every revision with the previous version and the next round's reports say what remains open.

To the editors and reviewers:

Please find hereby the revised version of our manuscript "FlexRL: an R package for Flexible Record Linkage and Linked Data Quality Assessment" which we resubmit for consideration for publication in the Journal of Reproducible Statistics. Please also find our answers to the comments of the review. We found these recommendations and points of discussion valuable, we incorporated them in the revised manuscript, with corresponding changes to the package code.

Please note that we have changed the random seeds since the first submission, having discovered that some of the compared packages call set.seed() internally. Thus, the results are slightly different from those previously reported, but we applied the same decision rules as before and revised our comments accordingly.

We hope that we have answered your concerns adequately and have improved the quality of the manuscript.

> REQUIRED Table 7 does not reconcile internally. Taking BRL, sensitivity from the table is TP/(TP+FN) = 80/(80+438) = 0.15, but the table reports 0.26; its FDP from FP/(TP+FP) = 21.90/(80+21.90) = 0.21, reported as 0.23. The same mismatch holds for multilink (sensitivity 137/517 = 0.27 vs 0.36; FDP 270.70/407.70 = 0.66 vs 0.58), fastLink, fedmatch and FlexRL. Please reconcile the reported rates with the counts, or state the aggregation rule in the caption. Tables 2, 3 and 6 reproduce their rates exactly from their counts, so Table 7 is the outlier.
>
> **Done.** The display of Table 7 was faulty. We now display the results on the same scale as Table 6 and reconciled the linkage performance over all regions.

> REQUIRED Table 6 and Table 7 appear to be on different scales. Table 6 (no blocking) implies 10568 true links (TP+FN, consistent across fastLink, fedmatch, FlexRL). Table 7 (with blocking) shows TP+FN of about 517 to 518 per method, roughly one twentieth of 10568, which matches a per-region average over the 20 regional blocks. Section 5 states region is stable and error free with agreement 1.00, so blocking on region should not drop true links from the national total. Please state whether Table 7 reports per-region averages or totals, since the two tables are presented as a direct comparison. A one-sentence caption note resolves it.

Done. Following the previous comment, we now present Table 7 on the same scale as Table 6 since the comparison between those holds importance in the section "Scalability". We now make it clear both tables report the national performance (with and without blocking).

REQUIRED No replications or Monte Carlo error are reported for the headline simulation tables. Table 2 and Table 4 appear to come from a single simulated pair of files (350/400 records). Several reported gaps are within plausible sampling noise of one run, for example FlexRL dynamics sensitivity 0.59 against fedmatch 0.60, and FlexRL FDP 0.16 against multilink 0.16. The FDP estimators are averaged over n_repeats, but the TP/FP/sensitivity/FDP comparison is not. Repeat the generate-and-link cycle across replications and report the spread of sensitivity and FDP, or qualify the dynamics claim to a single illustrative run and drop any ranking of methods that are within one run of each other.

Done. The simulation study presented in Table 2 and Tables 4 now report the mean and standard deviation over n_sim = 10 procedures, in each procedure the FDP estimation is based on n_repeats = 5 repetitions.

REQUIRED Add a competing-interests statement. The authors are the developers of FlexRL (Robach and Hof 2026) and the paper benchmarks that package against competing methods in Tables 2, 3, 6 and 7, so a disclosure of this role is a standard integrity obligation that is currently absent.

Done. We added such competing interests statement at the end.

REQUIRED In the SHIW application the score-based FDP estimate at the chosen threshold is about 0.03 while the true FDP is 0.16. State this gap explicitly where the print output appears, and note the synthetic estimate (0.12) was closer, so a practitioner is warned which estimator held up here.

Done. We added such comment in the "Diagnostics" section, as well as guidance on which estimator of the FDP to rely on in the "False Discovery proportion" section.

REQUIRED Document how random draws are produced inside the OpenMP-parallelised Rcpp samplers and whether set.seed() reproduces a run bit-for-bit when threads > 1. If it does not, qualify the reproducibility claim and state the thread count used for the reported tables.

Done. We documented the reproducibility with regard to the OpenMP parallelisation in section "Computational aspects". Indeed set.seed() reproduces a run bit-for-bit regardless of the number of threads.

REQUIRED For each benchmark table (2, 3, 6, 7), record the seed(s), R version, OS and hardware, and the version of every compared package, or state these in the text. The comparative timing, memory and performance numbers depend on all of them.

Done. Throughout the article (the quarto document) we (re)set the seed to 1 to produce each table and figure. We do so because some of the packages we compare FlexRL with call set.seed() within their source code, disturbing the random number generation. At the end of the "Simulation study" section, we now give these details as well as information on the R version, OS and hardware. The versions of the compared packages are given in Table 1.

RECOMMENDED The simulation DGP is the package's own simulate_data() function, which synthesises records from the same hierarchical generative model FlexRL fits (registration errors, missingness, a dynamic PIV governed by a hazard). This is a DGP aligned with the method under test, and the comparison-vector competitors are evaluated on data generated under assumptions they do not share. The headline claim that "modelling PIVs dynamics improves record linkage" rests on exactly this self-generated setting (Table 2: FlexRL dynamics sensitivity 0.59 vs no-dynamics 0.49). Report the same comparison on data generated outside the FlexRL model, or state in Section 2.4 that the simulation is a demonstration of the model's own DGP and that the dynamics benefit is quantified on real linkage only.

Done. The simulation study presented in Table 2 and Table 4 demonstrate the use of the simulate_data() function of the package and the potential of the FlexRL package (record linkage, diagnostic tools). The DGP is classical in record linkage and indeed aligns with the model FlexRL fits, we have clarified this in the "Simulation study" section. All methods are compared on the same simulated data. Moreover, to make a fair comparison of FlexRL record linkage with and without accounting for PIV dynamics, we generate the temporal changes with a piecewise constant hazard and fit these dynamics with an exponential survival model. This shows the benefits of modelling PIV dynamics rather than treating them as registration errors (even under model misspecification). We made this clear in the "Simulation study" section.

RECOMMENDED The simulation factor grid is a single configuration: one sample size, one overlap (300 of 350/400), 2% mistakes, 0.5% missingness, four PIVs, one dynamic PIV. The conditions that would stress the method and expose its failure modes, lower overlap, higher error rates, fewer discriminative PIVs, more dynamic PIVs, are not varied in the simulation. The real-data sections add scale and a national/regional contrast, but not a controlled sweep of these factors. Vary overlap and error rate across a small grid and report how the dynamics advantage and the FDP estimates hold up.

Declined. It is not the scope of this software article to study the behaviour of the compared packages over plausible record linkage configurations. In the "False discovery proportion" section, we refer the reader to the literature for a more detailed discussion on the the strengths and limitations of different record linkage paradigms across a range of configuration (including comparisons of FlexRL and BRL packages).

RECOMMENDED The threshold ξ=0.82 is chosen in Section 4.4 from the divergence "elbow", and in Section 4.1/ Figure 5 that choice is then shown to give a regression estimate close to the known truth. In the SHIW the truth is available, so this is a validation of the elbow heuristic, but the heuristic itself is the only guidance a practitioner has when the truth is unknown. State explicitly that the elbow rule is a heuristic and that its reliability away from benchmark data where true links exist has not been characterised.

Done. In the "Diagnostics" section we made it clear that the "elbow" of divergence measures is only a heuristic to choose the threshold and carries no formal guarantee.

RECOMMENDED The cross-method tables (Table 2, Table 3, Table 6, Table 7) compare methods at the single default rule ξ=0.5. Methods with different score scales and operating points are not necessarily comparable at one fixed cutoff, as the wide spread in sensitivity and FDP shows (for example reclin2 at sensitivity 0.80 with FDP 1.00 versus FlexRL at 0.33 and 0.26 in Table 3). Note in the table captions or surrounding text that these are single-threshold snapshots and that the FDP-versus-threshold curves give the fuller comparison.

Done. In the captions of the tables comparing record linkage methods we clarified the use of a single threshold. In the section "Linkage scores and decision rule" we mention that the "FDP versus linkage decision rule" curves provide a more comprehensive assessment of record linkage performance.

RECOMMENDED The Discussion describes FlexRL as a "reliable solution to record linkage". On the SHIW the method reaches sensitivity near 0.33 to 0.39 with FDP near 0.26 to 0.57, which reflects the difficulty of the task rather than a fault, but the wording overstates what the numbers show. Soften to describe the demonstrated behaviour, for example that the method controls FDP at a stated threshold while recovering a known fraction of links.

Done. We softened the "Discussion" with regard to the reliability of FlexRL record linkage and clarified the demonstrated behaviour: scalable in terms of memory and running time, flexible in terms of accommodating diverse scenarios, with performance depending on the difficulty of the task and the chosen decision rule, controlling the FDP and recovering a fraction of the true links (sensitivity) accordingly.

RECOMMENDED The second stated contribution is a method-agnostic toolkit of linked-data quality diagnostics. Edwards et al. 2026 ("Assessing Linkage Quality: Metrics, Insights, and a Reusable Python Toolkit", IJPDS) addresses the same problem of quality metrics for linked data in a reusable toolkit and is not cited. This is a live, actively-published line (the manuscript already cites Laidler 2025, Lam 2024, Harron 2017/2020 in the same venue). Please position the diagnostics contribution against it. The published metrics are the point here; comparison of the software itself sits with another assessor.

Declined. The details of the article "Assessing Linkage Quality: Metrics, Insights, and a Reusable Python Toolkit" (Edwards et al., IJPDS, 2026) do not yet seem available. From the main author website, this article appears to be a conference contribution. At this point, we therefore cannot position the diagnostics contribution of FlexRL against those provided by this toolkit.

RECOMMENDED The DESCRIPTION declares BRL, arf, diyar, fastLink, fedmatch, mice, multilink, reclin2 and synthpop as hard `Imports`. These are needed only by the optional `link_with_<pkg>()` wrappers and the synthetic-data FDP estimator, so loading FlexRL for its own StEM linkage pulls in nine extra packages, and a single install failure among them blocks FlexRL. Moving the wrapper and synthesis dependencies to `Suggests` with `requireNamespace()` guards at the call sites would lighten the install and match how wrappers are used.

Done. We modified the package accordingly to lighten its installation.

RECOMMENDED The log-likelihood sanity check in `simulateD()` compares `round(Dsample$loglik, 3) != round(ll_check, 3)` and calls `stop()` on mismatch. Equality after rounding to three decimals can false-trigger on large-magnitude log-likelihoods where a benign floating-point difference exceeds 0.0005, and can mask a real discrepancy below that scale. A relative tolerance (for example `isTRUE(all.equal(..., tolerance = 1e-6))`) would be more robust.

Done. We modified the package accordingly for a more robust log likelihood sanity check.

RECOMMENDED The FDP estimates in the print output and tables are given as point values (for example 0.03 at threshold 0.82 against a true FDP of 0.16). Since the synthetic-data estimator is already averaged over n_repeats draws, reporting the dispersion across those repeats would show how much of the gap between estimated and true FDP is sampling variability of the estimator.

Done. We modified the package to provide in the print output the mean and standard deviation of the FDP estimates obtained over n_repeats procedures. To lighten the "FDP versus linkage decision rule" figure we do not show the sample dispersion but the user can now access it in the output of the RL_diagnostics() function.

RECOMMENDED Figure 5 distinguishes four regression fits by colour alone: the text names "estimate in dashed orange", "estimate in yellow" and "truth represented in red". Orange and red read as close hues under the common red-green deficiencies, and yellow lines carry low contrast on a white ground. Redraw Figure 5 with a colourblind-safe palette and pair each line with a distinct line style or a direct label, so the three thresholds (0.55, 0.82, 0.99) and the truth can be told apart without relying on colour.

Done. We modified Figure 5 so it is interpretable without relying on colours.

RECOMMENDED One sentence in Section 3.5 reads "so the sweep is linear the number of candidate pairs", which is missing a word ("linear in the number"). The table column heading "FlexRL (no dynamic)" in Tables 2 and 4 would read better as "(no dynamics)" to match the "dynamics" wording used elsewhere.

Done. We modified the article accordingly.

RECOMMENDED State the provenance and access terms of the SHIW data in one sentence. The 2012 and 2014 waves are anonymised microdata released by the Bank of Italy for research, which places this work within secondary analysis of public data. Recording that status makes clear why separate ethics approval and participant consent are not required here.

Done. We elaborated on the SHIW provenance and access terms in the last paragraph of the "Introduction".

RECOMMENDED Qualify the word "reliable" in the Discussion given the reported FDPs (0.57 on the full data, 0.29 with blocking); say that the scalability claim concerns compute and memory, not linkage accuracy.

Done. Following a previous comment, we clarified the "Discussion" to position the scalability in terms of memory and running time and the flexibility in terms of accommodating diverse scenarios, with performance depending on the difficulty of the task and the chosen decision rule, controlling the FDP and recovering a fraction of the true links (sensitivity) accordingly.

RECOMMENDED Report the regression specifics: which SHIW variable encodes expected age of retirement, the linked sample size used (273 at the chosen threshold), the sensitivity there (0.26), and that the reference line is the estimate computed on true links.

Done. We reported the regression specifics in the "Diagnostics" section.

RECOMMENDED State how the PIV independence violation (Occupation and Sex associated, Cramér's V 0.32 and 0.35) may affect calibration of the FDP estimators, beyond issuing a warning.

Done. At the end of the section "False discovery proportion" we now elaborate on the assumptions underlying each FDP estimator and on the expected impact of deviations from the PIV independence assumption. We refer the reader to the FDP estimation article for a broader overview of the robustness of these FDP estimators to record linkage configurations.

RECOMMENDED Give practitioners guidance on which FDP estimator to trust when the two disagree, since the SHIW example shows them diverging from each other and from the truth.

Done. At the end of the section "False discovery proportion" we now provide guidance on which FDP estimator to rely on depending on the setting.

RECOMMENDED Clarify the scaling factor N_B/N_synth in Equation 11 and state when the synthetic estimator is expected to be biased. The simulation returns 0.28 against a true 0.16 for the dynamics model, which the text does not address.

Done. In the "Synthetic data estimator" section we gave details on the assumptions underlying the synthetic data FDP estimator and on its derivation. We commented on the performance of the estimators around Table 4.

RECOMMENDED Note that the linked-sample divergence measures are heuristic indicators of sampling bias, not estimates of the bias in β, and that the "elbow" threshold choice carries no formal guarantee.

Done. Following a previous comment, we clarified that the divergence measures are only heuristics and therefore neither estimate the sampling bias nor carry any formal guarantee.

RECOMMENDED Provide a single master script that maps each table and figure to the command that regenerates it, with a captured session-info / environment record.

Done. We added a preview of our session info at the beginning of the script. The quarto script is provided as a single document to generate the article and its content. We mentioned that the script is available as supplementary material and on the GitHub repository of the package at the end of the "Simulation study" section.

OPTIONAL The `.onAttach` startup message ("please cite us … feel free to complain as well") and the `music_on` argument that opens a browser tune are informal side effects. They are harmless and default off, but a plainer startup message reads better in a statistical-agency setting and avoids friction with package-check gates that flag chatty attach messages.

Declined. The message printing on attachment of the package can be silenced with suppressPackageStartupMessages() and does not interferes with the package checks. The music_on argument of the StEM() record linkage function is off by default and is convenient for long runs on very large data sets. We did not modify these elements.

OPTIONAL A broader simulation varying overlap, PIV discriminative power and error rates would map out when each FDP estimator is reliable.

Declined. It is not the scope of this software article to study the behaviour of FDP estimators across a range of record linkage configuration. However, such broader simulation study has been performed in the FDP estimation article and we refer the reader to it in the "False discovery proportion" section.

Round 2: Minor revision

SIGNED DECISION

Confirmed and signed by Matthias Templ on 2026-10-26 · ORCID 0000-0002-8638-5276

The three reviewers recommend acceptance and every required point of the first round is met. A minor revision is asked for the recommended points. The authors may leave any of them with a reason, and that will not hold the paper back.

DECISION LETTER

Thank you for the revised version of your manuscript. This is round 2, and we are inviting a further minor revision. Because you were asked for a major revision, this round read the new version whole, and the editor has read your response against it. What follows is our reading of this version. An item carried over from the previous round says so.

FlexRL packages a hierarchical two-source record linkage model (StEM) together with method-agnostic diagnostics, two FDP estimators and several sample-divergence measures, so a practitioner linking sources without a unique identifier can judge how the linkage decision rule affects downstream inference. It matters to statistical agencies and to researchers working with administrative, registry and survey data. We would be glad to see the revised version.

How to read this letter. The decision rests on the items marked required. If a required item seems wrong to you, or cannot be done within this paper, decline it on the revision page and give the reason. A reasoned decline is a legitimate answer, and the editor weighs it. Reviewers can be mistaken. Items marked recommended or optional are advice from the reviewers. Adopt them, decline them with a sentence, or leave them for later work. None of that counts against the paper. The reviewer reports below carry the same marks. The editor who signs this decision may weigh an item differently in the accompanying message, and that message prevails.

RECOMMENDED:

- [novelty-and-prior-art] The nearest uncited close work is Edwards et al. 2026, "Assessing Linkage Quality: Metrics, Insights, and a Reusable Python Toolkit" (IJPDS, doi:10.23889/ijpds.v11i5.3597), which addresses the same problem the paper claims as its second contribution: a method-agnostic toolkit of linked-data quality diagnostics. This was raised in round 1 (918f2fc9) and declined because the article was not yet available. The search now returns it as published in OpenAlex, so the ground for declining has changed. The two toolkits differ (FlexRL offers two FDP estimators and divergence measures rather than match-rate/precision/recall-style metrics), so this is positioning, not a claim collision. Cite it and state in the diagnostics section what the FDP and divergence tools add relative to the metrics Edwards et al. report.
- [statistical-rigor] The SHIW benchmark tables (3, 6, 7) report FlexRL from a single seed, yet FlexRL linkage is stochastic (StEM with an inner Gibbs sampler), so its counts carry Monte Carlo variability that the simulation tables now quantify but the real-data tables do not. Report the spread of FlexRL sensitivity and FDP across a small set of seeds on the Emilia-Romagna subset, or state in Section 5 that the reported real-data figures are a single realisation at seed 1 and note the dependence on it.
- [visualisation] The text uses "dotted orange" for the linked-data estimate at three different thresholds (0.55, 0.99 and 0.94) in Figure 5. If these appear together on one panel, add the threshold value as a direct label on each line so the three dotted lines are not confused. If they sit on separate panels, a one-line caption note saying so removes the ambiguity.
- [visualisation] Figure 3 shows the estimated FDP and several divergence measures against the threshold on shared axes. Confirm each curve carries a units label and that the FDP axis runs on a fixed 0 to 1 scale, so the elbow and the FDP level can be read without a truncated-axis effect. A direct label or legend entry per measure helps, since MMD, SMD and support IoU are on different scales.
- [software-landscape] The score-based FDP estimator (Equation 10) is the literature estimator used, among others, inside fastLink (Enamorado et al. 2019), yet the paper reports it only on FlexRL's own scores. A short line stating that Equation 10 is the general score estimator applied uniformly across all methods in Table 4, with one row showing the estimate FlexRL's wrapper computes from fastLink's own scores next to the FDP fastLink's native output would imply, would make clear the estimator is the published one and that the wrapper feeds it correctly.
- [software-quality] `cran-comments.md` states "an update from version 1.0.0 to 1.0.1" while `DESCRIPTION` is `Version: 2.0.0`, and README "News" still says "Version 1.0.0 of FlexRL is ready" and "will be soon available on JASP" though the paper describes FlexRL 2.0.0 already shipping as a JASP module. Refresh these before the CRAN/release submission so the metadata agrees with the version the paper documents.
- [Reviewer 2] In the SHIW example, note that the synthetic FDP estimate (0.07) carries a dispersion (sd 0.08) comparable to its own size, and that the near-zero dispersion of the score based estimate (0.01, sd 0.00) reflects confidence in a misspecified model, not precision about the true FDP of 0.15. New, visible because dispersions are now reported.

OPTIONAL:

- [simulation-design] The factor grid is still a single configuration, one overlap, one error rate, one dynamic PIV (prior point af410fe5, declined). The authors decline on scope grounds and point to the published configuration sweeps. A small grid varying overlap and error rate would show how the dynamics advantage and the FDP estimates hold up across difficulty, and could be parked for a later paper.

We will correct these in production, and you confirm them in the proof:

- [methods-soundness] The reference "Robach, Kayané, Mark A van de Wiel, Michel H Hof, and Stéphanie L van der Pas. 2026. Causal Record Linkage. h." still ends in a truncated URL, and the Lam et al. 2024 entry still carries a concatenated DOI ("v9i1.238910.23889/ijpds.v7i3.1794"). Confirm both links at proof.

REVIEWS

Report Reviewer 1, recommends accept

Written by the journal's review engine in the role of the application reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

I read the whole revision. Of the one required point and five lighter ones I raised, all are met in substance, and the one item still open is the incomplete reference, which is a production fix rather than an author task.

The required point was the honesty of the headline diagnostic in the worked example. The re-seeded run now lands the threshold at 0.94, and Section 4.4 states the gap directly: the synthetic data estimator yields 0.07 against a true FDP of 0.15, while the score based estimator yields 0.01. A reader now sees that the primary score estimator understated the error by an order of magnitude here and that the synthetic estimator held up, which is exactly the warning a practitioner needs. Section 4.2.2 goes further and explains why: under the PIV independence violation present in this example, the score based estimator inherits the model's misspecification, while the synthetic estimator stays valid if the synthesiser captures the PIV dependence. That reasoning is sound and it closes the point I was most concerned about, together with the related asks on the independence violation and on which estimator to trust when the two disagree.

The regression now names the SHIW variable ('etapen'), reports 245 retained links with sensitivity 0.23 and FDP 0.15 at the chosen threshold, and states that the solid reference line in Figure 5 is computed on the 898 true links. The Discussion no longer calls the method a "reliable solution" and instead separates the memory-and-runtime claim from linkage accuracy, which matches the FDPs in Tables 6 and 7. The simulation tables now report mean and standard deviation over 10 replications, and Tables 6 and 7 are on a common national scale, which makes the scalability comparison legible. These address the reproducibility and clarity points cleanly.

Two recommended points and two optional points were declined. The decline on not benchmarking the compared packages across a range of linkage configurations is reasonable for a software article, and the pointer to the methodological papers that do carry out that study is the right place for it. The decline on positioning against the Edwards et al. toolkit is also reasonable, since that work is not yet available to compare against. The optional declines on startup messages and on a broader FDP-estimator simulation do not affect the paper's claims. I am convinced by all four.

One item remains. The bibliography still shows "Robach, Kayané, Mark A van de Wiel, Michel H Hof, and Stéphanie L van der Pas. 2026. Causal Record Linkage. h." with a truncated venue field. This is cited for the inference consequences of linkage error, so it should carry a resolvable entry. It belongs on the production list the authors confirm in the proof.

Numbered suggestions to the authors:

1. [editor_fix] Complete the bibliography entry "Robach et al. 2026. Causal Record Linkage. h." with its venue or repository details.

Report Reviewer 2, recommends accept

Written by the journal's review engine in the role of the methodology reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

Of the required calibration point and the methodological clarifications I raised, the substantive ones are met. The paper now states in Section 4.4 that the score based estimator reads 0.01 against a true FDP of 0.15 on the SHIW while the synthetic estimator reads 0.07, and Section 4.2.2 explains the gap: the Sex and Occupation association the package itself flags violates the PIV independence that the score based estimator rests on, so that estimator understates the FDP, and the synthetic estimator should remain valid when the synthesiser captures the dependence. This is the honest account the headline tool needed, and the simulation in Table 4, where the PIVs are independent and the score based estimate (0.18 against 0.18

for the no-dynamics model) is well pinned, supports the mechanism. The decomposition in Equations 6 to 9, the StEM scheme in Algorithm 1 and the synthetic estimator derivation in Equation 11 are coherent as presented.

The seed change propagated new numbers throughout, and they hang together: the regional subset carries 898 true links, Table 3 gives FlexRL 298 true and 101 false positives at the default rule (FDP 0.25), and the ξ=0.94 analysis retains 245 pairs at FDP 0.15 and sensitivity 0.23. Table 4 now reports means and standard deviations over the replications, the print output carries the per-estimator spread, and the Table 4 text labels the synthetic estimate as unreliable for FlexRL (dynamics) and attributes it to the misspecified dynamics model. The divergence measures are now labelled as indicators of sampling bias that do not estimate Equation 9, and the elbow rule is stated as a heuristic whose reliability has not been assessed. These settle the points I raised.

One point is newly visible because the dispersions are now on the page, and it bears on the tool's practical use. In the SHIW example the synthetic estimate is 0.07 with a standard deviation of 0.08, so it is compatible with an FDP anywhere from near zero to roughly 0.23, and the true 0.15 sits inside that range. The score based estimate is 0.01 with a reported standard deviation of 0.00. A user choosing a threshold from the estimated FDP curve could read that 0.00 as precision when it is confidence in a model whose independence assumption is violated here. A sentence where the print output appears, noting the width of the synthetic estimate and what the near-zero score based spread does and does not mean, would let the reader weigh the two estimates as the authors intend. This strengthens the example and the paper stands without it.

The reference-level fixes I marked for production last round are still present: the truncated "h." URL on Robach et al. 2026, the concatenated DOI on Lam et al. 2024, and the JASP version that reads 1.0 in the text and 0.98.8 in the reference. These are confirmed in the proof rather than reworked by the authors.

The declined items are reasonably argued. A broader study of the compared packages and of the FDP estimators across linkage configurations sits outside a software paper and is referred to the methodological articles, the Edwards et al. toolkit is not yet available to position against, and the startup-message points are cosmetic. I am convinced on each.

Numbered suggestions to the authors:

1. [editor_fix] Complete the truncated "Robach et al. 2026, Causal Record Linkage" reference (URL still reads "h."), split the concatenated DOI in Lam et al. 2024, and reconcile the JASP version (text says 1.0, reference says 0.98.8).
2. RECOMMENDED Where the print output in Figure 4 is discussed, note that the synthetic FDP estimate (0.07) has a dispersion (sd 0.08) comparable to its own size, and that the near-zero dispersion of the score based estimate (0.01, sd 0.00) reflects confidence in a model whose independence assumption is violated here, not precision about the true FDP of 0.15.

Report Reviewer 3, recommends accept

Written by the journal's review engine in the role of the software reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

Of the two required points and one recommended point I raised, both required points are met and the recommended one is done, leaving a single production fix.

The reproducibility question I pressed on is now resolved in a way I find convincing. Section 3.5 states that all random draws occur upstream of the parallel region, so the OpenMP threads never call R's generator and only compute likelihood contributions over disjoint sets of candidate pairs, while the create-or-break proposals run as a sequential sweep. That design makes the output independent of the thread count, and set.seed() then fixes the whole run. This is the right answer to the concern, because it removes the thread-dependence at its source instead of asserting reproducibility and leaving the reader to guess what it covers. I could not run the samplers to confirm bit-for-bit identity across thread counts, so my confidence rests on the account rather than on execution, but the account is internally coherent and matches the stated architecture.

The provenance of the benchmark tables is now recorded. The hardware (MacBook Pro, Apple M1 Max, 32GB, 8 cores), R 4.6.1, the seed of 1 per output-producing cell, and the compared-package versions in

Table 1 are all present, and the note that some compared packages call set.seed() internally explains why the reported numbers shifted from the first submission. That disclosure is the useful part: a reader who regenerates the tables and sees different figures now knows where to look. The quarto document plus the session-info preview and the GitHub reproduction path cover the master-script request.

On the two declined points, both reasoned declines convince me. A study of how the competing packages behave across a range of linkage configurations, and a broader simulation of the FDP estimators, both sit outside a software paper and are already covered by the cited methodological articles. Deferring to that literature is the correct scope call here.

Reading the revised version as a whole, the changes to Tables 2 and 4 (mean and standard deviation over 10 replications, FDP averaged over 5 repeats) and the rescaling of Table 7 onto Table 6′s scale improve the paper without opening anything new that I would hold it for. The one item still outstanding is the truncated reference, which the authors' own bibliography determines and which belongs on the production list.

Numbered suggestions to the authors:

1. [editor_fix] Repair the reference "Robach et al. 2026. Causal Record Linkage. h." with its complete citation and URL.

REVISION, RECEIVED 2026-10-26

Authors confirm each required point as done or give their reason for declining it. Further answers are voluntary, because the review compares every revision with the previous version and the next round's reports say what remains open.

We made minor edits to the text to clarify these points.

> RECOMMENDED The nearest uncited close work is Edwards et al. 2026, "Assessing Linkage Quality: Metrics, Insights, and a Reusable Python Toolkit" (IJPDS, doi:10.23889/ijpds.v11i5.3597), which addresses the same problem the paper claims as its second contribution: a method-agnostic toolkit of linked-data quality diagnostics. This was raised in round 1 (918f2fc9) and declined because the article was not yet available. The search now returns it as published in OpenAlex, so the ground for declining has changed. The two toolkits differ (FlexRL offers two FDP estimators and divergence measures rather than match-rate/precision/recall-style metrics), so this is positioning, not a claim collision. Cite it and state in the diagnostics section what the FDP and divergence tools add relative to the metrics Edwards et al. report.
>
> **Declined.** We can only find an abstract of Edwards et al. research work and cannot find any Python code. For that reason we cannot examine the tools provided and cannot position them against the diagnostics provided in FlexRL.

> RECOMMENDED The SHIW benchmark tables (3, 6, 7) report FlexRL from a single seed, yet FlexRL linkage is stochastic (StEM with an inner Gibbs sampler), so its counts carry Monte Carlo variability that the simulation tables now quantify but the real-data tables do not. Report the spread of FlexRL sensitivity and FDP across a small set of seeds on the Emilia-Romagna subset, or state in Section 5 that the reported real-data figures are a single realisation at seed 1 and note the dependence on it.
>
> **Done.** Around Table 3, we stated that all results based on the SHIW data depend on the seed set in our script and on the seeds that some of the compared methods use internally.

> RECOMMENDED The text uses "dotted orange" for the linked-data estimate at three different thresholds (0.55, 0.99 and 0.94) in Figure 5. If these appear together on one panel, add the threshold value as a direct label on each line so the three dotted lines are not confused. If they sit on separate panels, a one-line caption note saying so removes the ambiguity.
>
> **Done.** The figures are arranged in separate panels, and the title of each panel informs the reader of the threshold (0.55, 0.94, 0.99). Each panel shows several curves, drawn in different colours and line styles, so they can be distinguished without relying on colour.

> RECOMMENDED Figure 3 shows the estimated FDP and several divergence measures against the threshold on shared axes. Confirm each curve carries a units label and that the FDP axis runs on a fixed 0 to 1 scale, so the elbow and the FDP level can be read without a truncated-axis effect. A direct label or legend entry per measure helps, since MMD, SMD and support IoU are on different scales.
>
> **Done.** Figure 3 presents multiple subfigures. All curve have their unit label and their axis. The FDP axis runs from 0 to 1 but in practice shows curves on the range 0 to 0.3. The elbow on the divergence

curves (namely MMD, SMD in A, SMD in B) and the FDP levels can be read without problem. Each curve is describe in the legend.

RECOMMENDED The score-based FDP estimator (Equation 10) is the literature estimator used, among others, inside fastLink (Enamorado et al. 2019), yet the paper reports it only on FlexRL's own scores. A short line stating that Equation 10 is the general score estimator applied uniformly across all methods in Table 4, with one row showing the estimate FlexRL's wrapper computes from fastLink's own scores next to the FDP fastLink's native output would imply, would make clear the estimator is the published one and that the wrapper feeds it correctly.

Done. In the "Score based estimator" section, we added references to the two articles that derive the score based FDP estimator of Equation 10. This simple derivation is implemented only in the FlexRL package and corresponds to the formula given in Enamorado et al. (2019) and Robach et al. (2025). In the "Diagnostics" section, we mention that diagnostics are run through wrappers with a common form for all compared methods.

RECOMMENDED `cran-comments.md` states "an update from version 1.0.0 to 1.0.1" while `DESCRIPTION` is `Version: 2.0.0`, and README "News" still says "Version 1.0.0 of FlexRL is ready" and "will be soon available on JASP" though the paper describes FlexRL 2.0.0 already shipping as a JASP module. Refresh these before the CRAN/release submission so the metadata agrees with the version the paper documents.

Done. We updated cran-comments.md, DESCRIPTION and NEWS.md accordingly. FlexRL has been prepared as a JASP module for JASP 1.0, which has not yet been released. We will submit FlexRL 2.0.0 to CRAN once the journal has approved all changes.

RECOMMENDED In the SHIW example, note that the synthetic FDP estimate (0.07) carries a dispersion (sd 0.08) comparable to its own size, and that the near-zero dispersion of the score based estimate (0.01, sd 0.00) reflects confidence in a misspecified model, not precision about the true FDP of 0.15. New, visible because dispersions are now reported.

Done. We softened the sentence to present both FDP estimators in a similar fashion. In the "False discovery proportion" section, we provide guidance on which estimator of the FDP to rely on depending on the setting.

OPTIONAL The factor grid is still a single configuration, one overlap, one error rate, one dynamic PIV (prior point af410fe5, declined). The authors decline on scope grounds and point to the published configuration sweeps. A small grid varying overlap and error rate would show how the dynamics advantage and the FDP estimates hold up across difficulty, and could be parked for a later paper.

Done. Both points have been investigated in articles we reference. The behaviour of the two FDP estimators implemented in FlexRL across configurations is studied in the article cited in the "False discovery proportion" section. The benefit of modelling PIV dynamics, compared for FlexRL with and without dynamics and with other record linkage paradigms, is studied in the article cited in the "Simulation study" section. We clarified these contributions.

Round 3: Accepted

SIGNED DECISION

Signed by Matthias Templ on 2026-10-28 · ORCID 0000-0002-8638-5276

Final acceptance of version 2, after the editor read the preview of the article. The journal's re-run reproduces the reported values on macOS arm64, except the BRL and multilink cells of Tables 2 and 4, which come from other packages' samplers, and fastLink in Table 6, which needs 20 GB of memory. One required point remains for the proof: the text names JASP 1.0 and the reference version 0.98.8.

DECISION LETTER

Congratulations, your manuscript has been accepted for publication in JRepStat. Production begins now.

Where it is not on file yet, we still need the following for production.

- a valid ORCID iD for every author
- a captured computational environment (dependencies/lockfile/container)

- a signed determinism attestation for the entry-point command

REVIEWS

Report Reviewer 1, recommends accept

Written by the journal's review engine in the role of the application reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

The one item I left open last round was the single editor_fix on the truncated reference, since everything else was met in substance. That item is still open. The bibliography still reads "Robach, Kayané, Mark A van de Wiel, Michel H Hof, and Stéphanie L van der Pas. 2026. Causal Record Linkage. h." with no resolvable venue or repository. It is cited for the inference consequences of linkage error, so it should carry a complete entry. This belongs on the production list the authors confirm in the proof.

Nothing else is open. My recommendation is unchanged.

Numbered suggestions to the authors:

1. [editor_fix] Complete the bibliography entry "Robach et al. 2026. Causal Record Linkage. h." with its venue or repository details.

Report Reviewer 2, recommends accept

Written by the journal's review engine in the role of the methodology reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

Of the two items I carried into the accept round, the recommended dispersion note is not in the text and the three reference fixes are unchanged. Neither bears on the recommendation, which stands at accept.

The reference-level items remain for the proof. The truncated "Causal Record Linkage" URL still reads "h.", the Lam et al. 2024 DOI is still two DOIs run together, and the text still says the JASP module is for version 1.0 while the reference cites version 0.98.8. The response letter explains the JASP point, that the module targets an unreleased JASP 1.0 while the current release is 0.98.8, which makes the two numbers defensible rather than contradictory. The editor should still decide whether to align them in the proof, so I leave the item on the production list rather than close it.

The recommended dispersion note was voluntary and the authors did not take it up. In the print output the synthetic estimate is 0.07 with a standard deviation of 0.08 and the score based estimate is 0.01 with a reported standard deviation of 0.00, against a true FDP of 0.15. A sentence reading those two spreads would help a user weigh the estimates, since the near-zero score based spread reflects confidence in a model whose independence assumption is violated here, not precision about the true value. The paper stands without it, and the Section 4.2.2 account of why the score based estimator understates the FDP already carries the main point.

Numbered suggestions to the authors:

1. [editor_fix] Complete the truncated "Robach et al. 2026, Causal Record Linkage" reference (URL still reads "h."), split the concatenated DOI in Lam et al. 2024, and reconcile the JASP version between the text (1.0) and the reference (0.98.8).
2. RECOMMENDED Where the print output in Figure 4 is discussed, add a sentence noting that the synthetic FDP estimate (0.07) has a dispersion (sd 0.08) comparable to its own size, and that the near-zero dispersion of the score based estimate (0.01, sd 0.00) reflects confidence in a model whose independence assumption is violated here, not precision about the true FDP of 0.15.

Report Reviewer 3, recommends accept

Written by the journal's review engine in the role of the software reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

My prior report left a single open item, the truncated "Causal Record Linkage" reference, and that is the one point to settle here.

It is still truncated. In the revised bibliography the entry reads "Robach, Kayané, Mark A van de Wiel, Michel H Hof, and Stéphanie L van der Pas. 2026. Causal Record Linkage. h.", where the trailing "h." is a cut-off URL. The response letter answered a set of recommended and optional points but did not touch

this one, so the error carries over unchanged. Its correct value is fixed by the authors' own citation, which keeps it on the production list rather than the Required list, and it does not bear on any claim in the paper.

Nothing else in this version changes my view. The reproducibility design I pressed on, all random draws upstream of the parallel region so thread count does not affect the output, and the benchmark provenance (hardware, R 4.6.1, seed of 1 per output cell, compared-package versions) both stand as in the round I reviewed. The recommendation remains accept, with the one reference repair handled in the proof.

Numbered suggestions to the authors:

1. [editor_fix] Repair the reference "Robach, Kayané, Mark A van de Wiel, Michel H Hof, and Stéphanie L van der Pas. 2026. Causal Record Linkage. h." with its complete citation and URL.