ATLAS: An automated association test using probabilistically linked health records with application to genetic studies

Harrison G. Zhang[1,2,3]*, Boris P. Hejblum [4,5]*, Griffin M. Weber [1], Nathan P. Palmer [1], Susanne E. Churchill [1], Peter Szolovits [6], Shawn N. Murphy [7,8], Katherine P. Liao [1,2], Isaac S. Kohane [1], Tianxi Cai [1,4]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA;

[2]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Boston, MA, USA;

[3]Department of Biological Sciences, Columbia University, New York City, NY, USA;

[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA;

[5]Univ. Bordeaux, ISPED, Bordeaux Population Health Research Center, Inserm U1219, Inria SISTM, 33000, Bordeaux, France;

[6]Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, MA, USA;

[7]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA;

[8]Research IS and Computing, Mass General Brigham HealthCare, Charlestown, MA, USA;

*These authors contributed equally.

Correspondence to: Tianxi Cai, 665 Huntington Avenue, Boston, MA 02115;

tcai@hsph.harvard.edu

Keywords: electronic health records; record linkage; genetic association studies; biorepositories; perturbation resampling.

Word Count: 3997

Abstract

Objective: Large amounts of health data are becoming available for biomedical research. Synthesizing information across databases with no gold standard mappings between records may provide a more complete picture of patient health and enable novel research studies. To do so, researchers may probabilistically link databases and conduct inference using the linked data. However, previous inference methods for linked data are constrained to specific linkage settings and exhibit low power. Here, we present ATLAS, an automated, flexible, and robust association testing algorithm for probabilistically linked data.

Materials and Methods: Missing variables are imputed at various thresholds using a weighted average method that propagates uncertainty from the linkage process. Next, an estimated effect size is obtained using a generalized linear model. ATLAS then conducts the threshold combination test by optimally combining p-values obtained from data imputed at varying thresholds using Fisher's method and perturbation resampling.

Results: In simulations, ATLAS controls for type I error and exhibits high power compared to previous methods. In a real-world application study, incorporation of linked data-enabled analyses using ATLAS yielded two additional signifigant associations between rheumatoid arthritis genetic risk score and biomarkers.

Discussion: The ATLAS weighted average imputation weathers false matches and increases contribution of true matches to mitigate linkage error induced bias. ATLAS' threshold combination test avoids arbitrarily choosing a threshold to rule a match, thus automating linked data-enabled analyses and preserving power.

Conclusion: ATLAS promises to enable novel and powerful research studies using linked data to capitalize on all available data sources.

1. Introduction

A vast amount of health data stemming from electronic health records (EHR), biorepositories, administrative claims, and biomedical research studies are becoming available for discovery and predictive research [1,2]. For patients who contribute information to multiple databases, synthesizing their information across all available sources captures a more complete picture of their health and allows for more comprehensive and powerful research studies. For example, database A may contain genomic data and database B may contain longitudinal phenotypic data. Linking patient records in these databases would allow researchers to investigate gene-disease associations. In a similar vein, researchers recently linked genomics data with environmental factors of BRCA1/BRCA2 mutation carriers from two independent studies to study environmental-gene relations, demonstrating the potential to conduct innovative research studies after linking databases [3].

To perform linkage when protected health information (PHI) identifiers are not available, as is generally the case in de-identified research databases, researchers employ probabilistic record linkage (PRL) [4]. However, linkage errors are inevitable in PRL due to data discrepancies, and examples of linkage errors include incorrectly linking two records that do not belong to the same patient (false matches) or leaving a record unlinked when a correct link exists (missed matches). Neter *et al.* were the first to investigate the consequences of linkage errors on downstream inference results, and they showed that such errors induce substantial bias in inference [5]. More recently, Rentsch *et al.* attempted to perform inference on linked real-world data and provided evidence that false matches reduce magnitudes of association while leading to highly biased estimates [6]. We are specifically interested in testing the association between some predictors X

– recorded in one database A, and an outcome Y – recorded in a different database B. Within an association testing framework, both false matches and missed matches drive the results in the direction of no association by undermining statistical power [7–9]. This is because false matches increase sample sizes but dilute potential associations while missed matches reduce sample sizes and undermine statistical efficiency [7–9].

Despite the rapid growth in analysis of linked data and the well documented effects of linkage error on downstream inference, few robust, automated, and flexible inference methods have been proposed to account for linkage error induced bias [10]. Further, current proposed estimators are restricted to specific linkage settings and few are implemented in open-source software. For example, Hof *et al.* propose to address uncertainty from PRL by weighting least square estimators in linear and logistic regression [11]. But, their model does not formally account for non-match events, and they conclude that their estimator is biased unless complete matching is achieved [11]. Chipperfield proposes a weighting approach to make inference about regression coefficients but assumes that database B must be a complete subset of database A [12]. Dalzell *et al.*'s inference method necessitates the selection of extraneous blocking variables and, they demonstrate that block structure is needed to make unbiased estimates [13]. Recently, Han *et al.* propose linkage bias correcting estimators that do not make specific assumptions about linkage settings or require specific data structures as previously proposed methods do [14]. However, their method does not account for linkage settings where covariates come from either database A or B [14].

In this article, we propose automated association testing using probabilistically linked health records (ATLAS), a fully automated and scalable association testing framework that addresses many of these limitations. ATLAS utilizes either (1) a best match method or (2) a weighted

imputation method that propagates uncertainty from the linkage process contained in matching probabilities. Then, ATLAS optimally combines several p-values estimated using generalized linear models (GLMs) that each correspond to a different matching threshold ρ_k ($k \in \{1, \dots, K\}$) as a significance test, avoiding the difficult choice of choosing a single threshold for defining a match. Unlike previous work, ATLAS performance is not conditional on specific linkage settings, linkage methods, or data structure. To facilitate ease of use and accessibility, we have implemented ATLAS as an open-source R package on CRAN. Here, we validate ATLAS performance and compare to existing methods in simulation and real-world studies to show that ATLAS is more robust at detecting associations than previously published estimators.

2. Materials and Methods

2.1. Statistical method

The proposed ATLAS algorithm broadly consists of 3 steps: (1) missing variable imputation using either i) the best match above a pre-specified threshold, or ii) a weighted average using matching probabilities as weights also filtering on a pre-specified threshold; (2) estimation of an adjusted effect size using a GLM; and (3) a significance test relying on optimal combination of p-values obtained from data imputed at varying thresholds using Fisher's method together with perturbation resampling. Figure 1 illustrates this procedure and each of these three steps is further detailed in following subsections. Without loss of generality and for the sake of simplicity, here we will consider the situation where we have database A containing a p-dimensional novel predictor vector X and a vector of matching features M_A on n_A subjects indexed by $i \in \{1, \dots, n_A\}$ and database B containing outcome information Y, a vector of matching features M_B, and potentially some other existing covariates W on n_B subjects indexed

by $j \in \{1, \dots, n_B\}$. We seek to link the predictor variables X recorded in database A subjects in database B such that we may run an association analysis of $Y \sim W + X$.

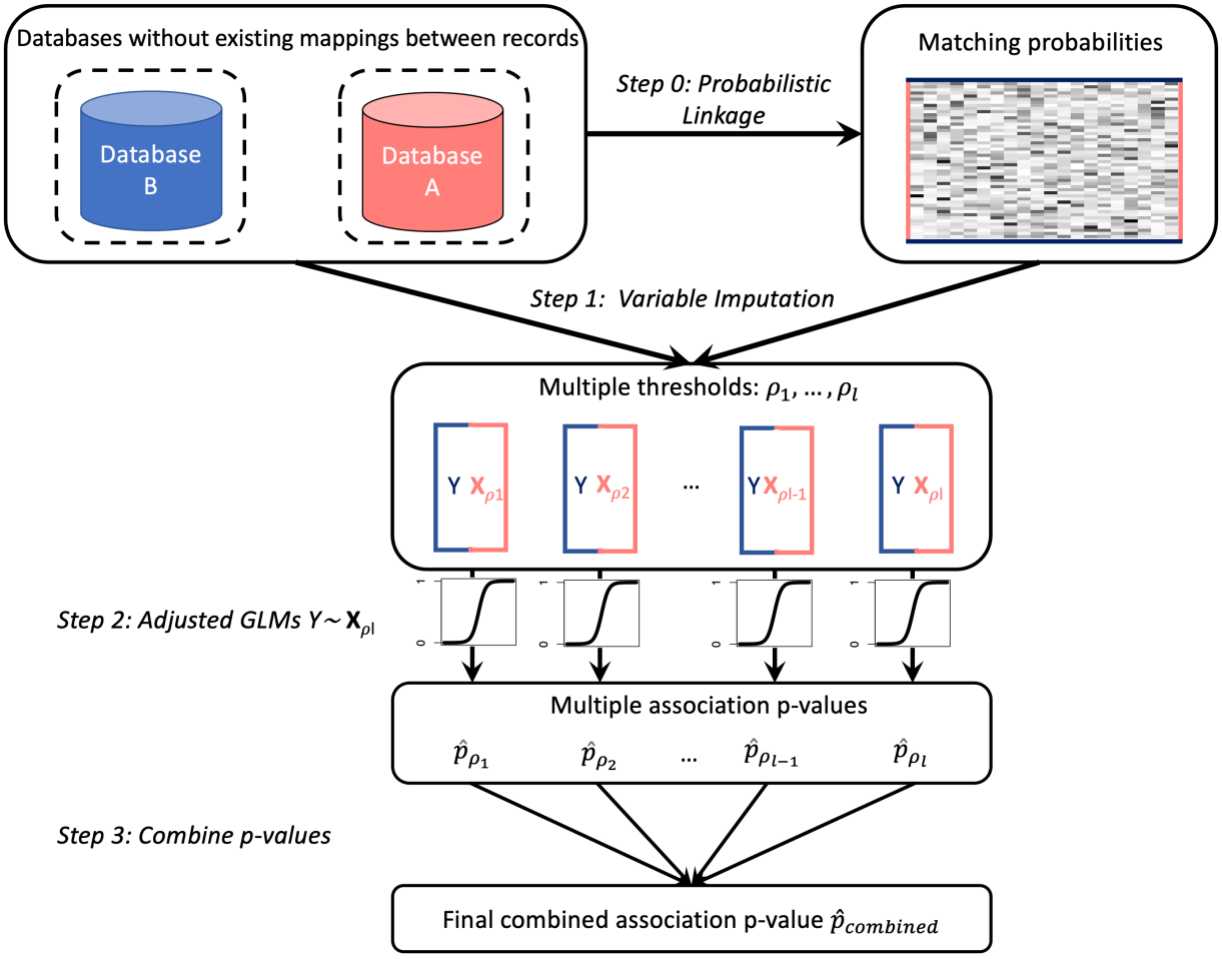

Figure 1: Schematic of the proposed ATLAS algorithm

2.1.1. Missing variable imputation

The linkage problem can be seen as a missing information problem. We denote the probability of matching between a patient i from database A and a patient j from database B as π_{ij}, where π_{ij} is ascertained via any given linkage algorithm that typically assess the similarity between the matching features M_{Ai} and M_{Bj}. For example, the `ludic` linkage algorithm employs Bayesian

modeling of binary diagnosis codes as matching features to estimate a posterior probability of being a match given a pair of patient records [15]. Using these estimated probabilities and given a specific matching cutoff threshold ρ_k above which patients are ruled as a match, we define $\xi_j(\rho)$ as whether there is any match for patient j in database B:

$$\xi_j(\rho) = \begin{cases} 1 & \text{if } \max\limits_{1 \le i \le n_A} \pi_{ij} \ge \rho \\ 0 & \text{else} \end{cases}$$

The best match imputation method imputes missing a Z_j with the observation X_i from patient i in A with the highest probability $\hat{\pi}_{ij}$ above a threshold ρ_k. However, this method risks imputing missing variables from false matches created by linkage errors and diluting potential associations. Therefore, we propose instead a weighted average imputation of missing variables using linkage probabilities as weights to weather false matches and increase contribution of true matches. More specifically, for the j^{th} subject in database B and a given threshold ρ, we identify subjects in database A with linkage probabilities $\hat{\pi}_{ij}$ above ρ and obtain a weighted average of their X's as the linked predictor:

$$\hat{X}_{\rho j} = \frac{\sum_{i=1}^{n_A} \pi_{ij}\, I(\pi_{ij} \ge \rho) X_i}{\sum_{i=1}^{n_A} \pi_{ij}\, I(\pi_{ij} \ge \rho)}, j = 1, \dots, n_B$$

If $\max_{1 \le i \le n_A} \pi_{ij} < \rho$, then the predictor will be deemed as missing for the j^{th} subject.

2.1.2. Effect size estimation using Generalized Linear Models (GLM)

Our goal is to assess the association between X and Y using the linked database B through a GLM with link function g:

$$E(Y|X, W) = g(W'\alpha_0 + X'\theta_0) = g(Z'\beta_0)$$

, where $Z = (W', X')'$ and $\beta = (\alpha', \theta')'$. For simplicity, we focus on the downstream association testing for $H_0: \theta = 0$ using linked data

$$\mathbb{D}_{\mathbb{B}}(\rho) = \left\{\left(Y_j, \hat{Z}_{\rho j}\right): \xi_j(\rho) = 1, j = 1, \dots, n_B\right\}$$

where $\hat{Z}_{\rho j} = \left(W'_j, \widehat{X}'_{\rho j}\right)'$. We fit the GLM $Y_j \sim \hat{Z}_{\rho j}$ to $\mathbb{D}_{\mathbb{B}}$ to obtain a maximum likelihood estimate for β, denoted as $\hat{\beta}_\rho = \left(\widehat{\alpha'}_\rho, \widehat{\theta'}_\rho\right)'$, and test for $H_0: \theta_0 = 0$ based on the corresponding $\hat{\theta}_\rho$.

2.1.3. The ATLAS threshold combination test: significance testing with optimal p-value combination

Various threshold values can be used for ρ. For instance, one could use $\rho = 0.5$, where indicated matches have a higher probability of being match than non-match, or $\rho = 0.9$, where indicated matches have higher certainty of being true matches. However, the optimal choice of such a threshold is unclear in practice due to the lack of gold standard labels on the true mappings between A and B. On the one hand, higher thresholds have lower estimation biases from fewer false matches but at the price of decreased statistical power from smaller sample sizes. On the other hand, lower thresholds exhibit higher statistical power at the expense of increased estimation bias.

Thus, instead of arbitrarily choosing a threshold ρ, we propose to optimally combine several p-values that correspond to different matching thresholds $\{\rho_l, l = 1, \dots, L\}$, thereby automating the significance testing process and preserving statistical power in various settings. Specifically, we propose to obtain a p-value via a χ^2 test based on $\hat{\theta}_{\rho_l}$ for the threshold ρ_l, $\hat{p}_{\rho_l} = P\left(\chi^2_p \geq \widehat{\theta'}_{\rho_l} \hat{\Sigma}^{-1}_{\rho_l} \hat{\theta}_{\rho_l}\right)$, and construct a combined test statistic as: $\hat{\gamma} = \sum_{l=1}^{L} -\log(\hat{p}_{\rho_l})$ to calculate the final

p-value for the testing of $H_0: \theta_0 = 0$. For a given ρ_l, we may obtain p-value $\hat{p}_{\rho_l} = P\left(\chi_p^2 \geq \widehat{\theta'}_\rho \hat{\Sigma}_\rho^{-1} \hat{\theta}_\rho\right)$ via a p-degree of freedom χ^2 test, where $\hat{\Sigma}_{\rho l}$ is the estimated variance-covariance matrix of $\hat{\theta}_{\rho l}$. Since $\left\{\hat{\theta}_{\rho_l}, l = 1, \ldots, L\right\}$ are estimated using overlapping data, the test statistics $\left\{\widehat{\theta'}_{\rho_l} \hat{\Sigma}_{\rho_l}^{-1} \hat{\theta}_{\rho_l}, l = 1, \ldots, L\right\}$ are highly correlated with each other. Thus, to estimate the null distribution of $\hat{\gamma}$, we use a resampling strategy to account for the correlations. Specifically, we generate a vector of n_B standard gaussian random variables $\mathbb{G} = \left(G_1, \ldots, G_{n_B}\right)'$, and subsequently obtain a perturbed random vector as $\hat{\theta}_\rho^{[\mathbb{G}]} = \sum_{j=1}^{n_B} \xi_j(\rho_l) \hat{S}_{\theta j}(\rho_l) G_j$, where

$$\hat{S}_j(\rho) = \left[\hat{S}_{\alpha j}(\rho)', \hat{S}_{\theta j}(\rho)'\right]' = J(\rho)^{-1} \hat{Z}_{\rho j} \{Y_j - g(\widehat{\beta'}_\rho \hat{Z}_{\rho j})\},$$

$J(\rho) = \sum_{j=1}^{n_B} \xi_j(\rho) \hat{Z}_{\rho j} \widehat{Z'}_{\rho j} \dot{g}(\widehat{\beta'}_\rho \hat{Z}_{\rho j})$ is the Fisher Information matrix for a given ρ and $\dot{g}(x) = \frac{dg(x)}{dx}$. Subsequently, we obtain the perturbed counterpart of $\hat{\gamma}$ as $\hat{\gamma}^{[\mathbb{G}]} = \sum_{l=1}^{L} -\log\left(\hat{p}_{\rho_l}^{[\mathbb{G}]}\right)$, where

$$\hat{p}_{\rho_l}^{[\mathbb{G}]} = P\left(\chi_p^2 \geq \hat{\theta}_\rho^{[\mathbb{G}]\prime} \hat{\Sigma}_\rho^{-1} \hat{\theta}_\rho^{[\mathbb{G}]}\right).$$

We then generate R realizations of $\mathbb{G}, \{\mathbb{G}_r, r = 1, \ldots, R\}$, to obtain the final p-value for testing $H_0: \theta_0 = 0$ as:

$$\hat{p}_{\text{combined}} = 1 - \frac{1}{R} \sum_{r=1}^{R} 1_{\{\hat{\gamma} \geq \hat{\gamma}^{[\mathbb{G}r]}\}}$$

2.2. Data and metrics for evaluation

We evaluated the performance of ATLAS in simulation studies and conducted a real-world genetic association study using EHR data that has been linked to a biorepository.

2.2.1. Simulation study

In our simulation studies, we estimated type I error rate (statistical size) as well as empirical statistical power of ATLAS under various settings using a publicly available de-identified database with $N = 2{,}723$ patient records containing 1,342 unique International Classification of Disease (ICD), Ninth Revision codes [15]. To test robustness of ATLAS under different linkage and inference settings, our simulations compared ATLAS performance by (1) creating discrepancies between databases by perturbing with multivariate noise, (2) adjusting the average codes per patient record, (3) adjusting effect sizes, and (4) using outputs from different linkage algorithms to test compatibility with ATLAS. We considered two linkage algorithms in our simulations: i) `ludic`, a published algorithm which relies on Bayesian modeling of binary diagnosis codes, and ii) `embeddingMatch`, which calculates cosine similarities between patient-level semantic embedding vectors (SEVs) similar to other embedding based linkage methods [15–18]. Further details regarding the `embeddingMatch` method and the simulation model are detailed in the Supplementary Appendix. Type I error rates using single cutoff thresholds and for the ATLAS threshold combination test were estimated as the proportion of p-values less than the nominal testing level $\alpha = 0.05$ under no simulated association. Similarly, empirical power was estimated as the proportion of significant p-values at $\alpha = 0.05$ under simulated association for a given $\mathrm{OR} > 1$. Results are based on $n = 1{,}500$ simulations in each setting, and thresholds used in the ATLAS threshold combination test include $\rho \in (0.1, 0.3, \ldots, 0.9)$. For the sake of simplicity, we report ATLAS results at single cuttoff thresholds of 0.1, 0.5, and 0.9, and report the rest in the Supplementary Appendix.

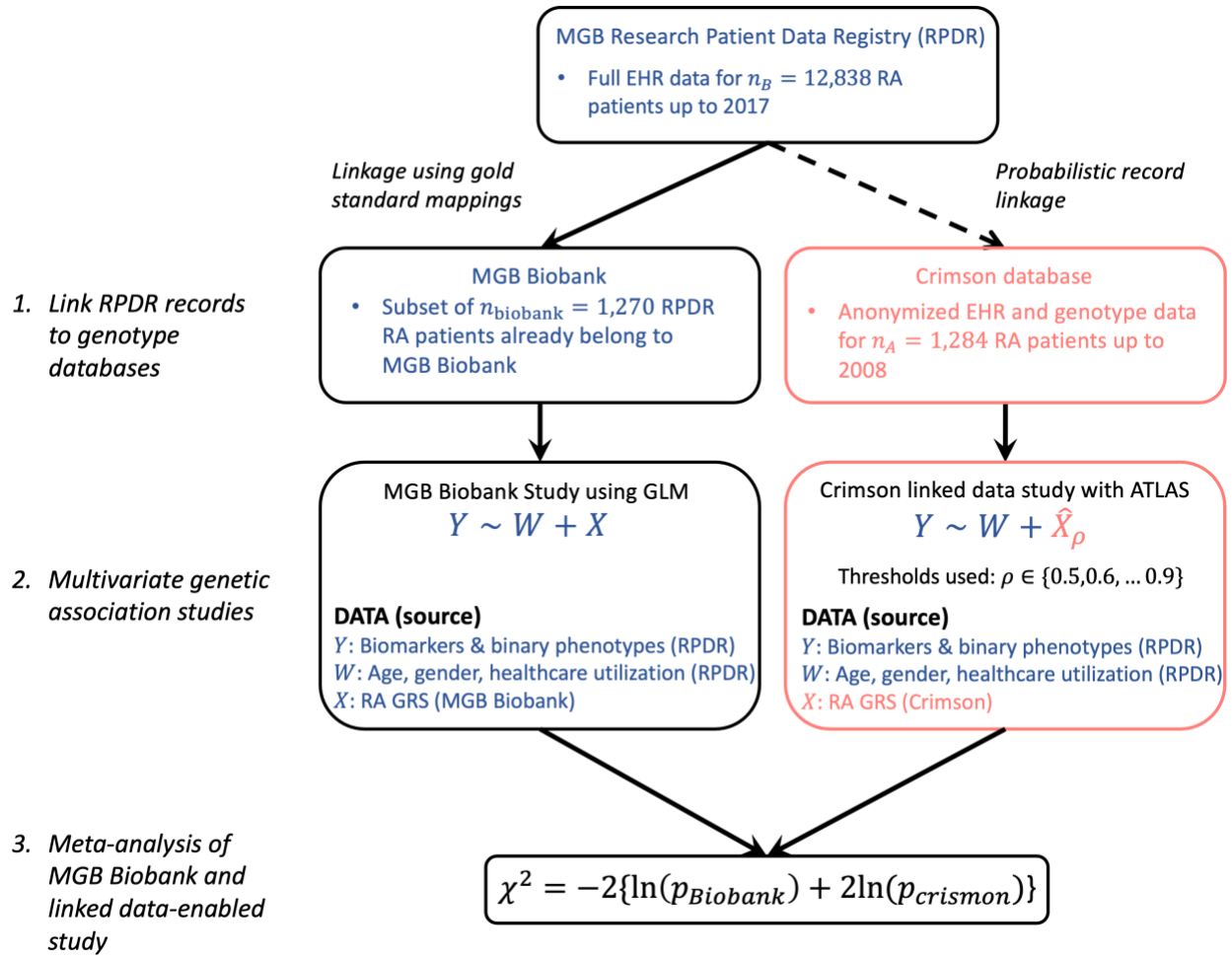

Figure 2: Schematic of real-world genetic association studies conducted using the MGB Biobank and the Crimson linked data.

2.2.2. Genetic association study using real-world biorepository data

To further validate performance and demonstrate real-world utility of ATLAS, we conducted a genetic association study with linked data to assess the association between rheumatoid arthritis (RA) genetic risk score (GRS) and various clinical outcomes among RA patients. We considered EHR records from two databases with which we performed linkage and the linked data-enabled downstream association analyses. The EHR data are stored in two separate databases: (i) the Crimson Clinical Discards database (herein referred to as Crimson and akin to database A) for a

subset of RA patients of European descent previously identified in 2008 [19,20]; and (ii) the Mass General Brigham (MGB) Research Patient Data Registry (herein referred to as RPDR and akin to database B) subset of RA patients identified via an existing machine learning algorithm [21–23]. The Crimson RA cohort contains anonymized EHR data along with genotype data for $n_{\text{Crimson}} = 1{,}284$ patients collected up to 2008. The RPDR RA cohort contains full EHR data up to 2017 for a total of $n_{\text{RPDR}} = 12{,}838$ patients. A subset of $n_{\text{biobank}} = 1{,}270$ RPDR patients already report genotype data because they belong to the MGB Biobank, and we have gold standard labels on the true mappings between RPDR and MGB Biobank records. [24,25]. Our goal is to link Crimson with RPDR to enable genetic association studies of RA GRS against various clinical outcomes using all available data across RA cohorts. Figure 2 provides a schematic of our analyses. Details about how RA GRS were constructed are reported in the Supplementary Appendix [20,26,27].

To perform record linkage, we assembled all available ICD codes in Crimson records and ICD codes recorded prior and up to 2008 in RPDR records. ICD codes were then aggregated to PheCodes using the standard procedure, and a vector of 1,542 binary matching features was created for each patient record where each feature was a binary indicator of presence or abscence of a PheCode [28,29]. Next, we performed PRL using `ludic` to estimate probabilities of being a match for every possible pair of RPDR and Crimson records. Since the Crimson cohort consists of RA patients that were previously managed at MGB, we anticipated that a majority of these patients can be linked to the updated RPDR RA cohort and that some of the linked patients may report genotype data from both MGB Biobank and Crimson. In the absence of gold standard labels on the true mappings between RPDR and Crimson records, we validate the accuracy of the

linkage by assessing the concordance between RA GRS from MGB Biobank and those from Crimson among the matched subset with genotype data available from both databases.

Once Crimson records are linked to RA RPDR records, we imputed GRS values using the weighted average method from Crimson to those subjects in RA RPDR using thresholds $\rho_l \in \{0.5, 0.6, \ldots, 0.9\}$. Subsequently, we used ATLAS to conduct multivariate association studies of linked RA GRS and clinical outcomes while adjusting for patient age, gender, and healthcare utilization (defined as the $\log[1 + \text{total encounters}]$). Clinical outcomes from RDPR include laboratory biomarkers commonly used to assess patient inflammation and binary phenotypes for pyogenic arthritis and gout, which are other distinct non-autoimmune forms of arthritis. These binary phenotypes were defined as having at least 2 PheCodes corresponding to these disorders and were constructed using ICD codes recorded up to 2017 in RPDR [28,29,29]. Reported effect sizes were estimated using data imputed at a $\rho = 0.9$ threshold.

Using RDPR patients who already belong to the MGB Biobank, we replicated these multivariate association studies and conducted meta-analysis using Fisher's method to combine the P-values estimated from the MGB Biobank and Crimson linked-data study to demonstrate increased power to detect associations when linking databases. To determine statistical signifiance after meta-analysis, we accounted for multiple testing by adjusting p-values to control for a false discovery rate (FDR) of 5% using the Benjamini-Hochberg procedure [30].

2.2.3. Benchmark methodology for comparison

In both our simulation studies and our real-world association study, we additionally considered the bias correcting estimators for linked data proposed by Han *et al.* as benchmark approaches [14]. The three Han *et al.* estimators are "Han F" (for "Full" – which considers all possible pairs

for each patient in *A*), "Han M" (for "Max" – which considers largest probabilities for each patient in *A*), and "Han M2" (for "Max 2" – which considers two largest probabilities for each patient in *A*) [14]. We report type I error rates and empirical power of these estimators. We were unable to replicate the real-world multivariate study using Han *et al.* estimators because they do not consider accepting covariates from the database that provides outcomes. To compare performance in the real-world setting, we conducted univariate analyses.

3. Results

3.1. Type I error control in simulations

In Figure 2 we show estimated type I error rates of ATLAS under linkage conditions simulated according to (1) average number of codes per patient record and (2) discrepancies between the databases created by perturbing with multivariate noise. ATLAS effectively controlled for type I error in all simulation settings and at all considered thresholds regardless of the linkage or downstream imputation method used. Han et al.'s estimators controlled for type I error, although they appear somewhat too conservative.

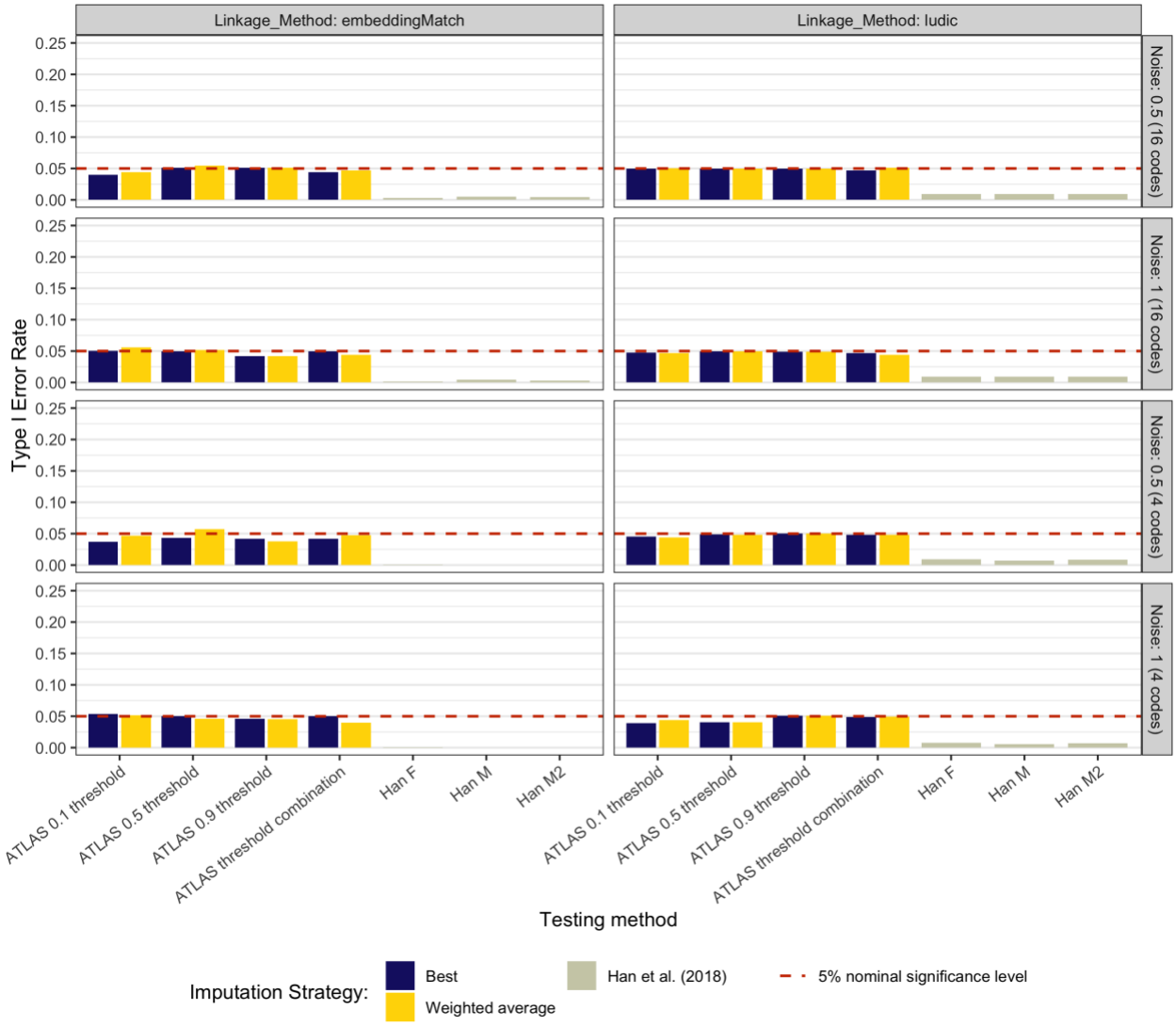

Figure 3: Comparison of type I error rates of ATLAS and Han et al. estimators in simulation settings with different noise levels and average codes per patient record (simulations under H_0). ATLAS type I error rates reported for several single cutoff thresholds and the ATLAS threshold combination test.

3.2. Statistical power evaluation

Figure 3 shows empirical power of ATLAS using best match imputation. Under linkage conditions where patient records had on average 16 codes, there was little difference in power at

different single-cutoff thresholds. Under conditions where patient records had on average 4 codes, we observed more significant differences in power between single cutoff thresholds. Relative power between thresholds was dependent on the linkage method. However, across all settings, power exhibited by the ATLAS threshold combination test demonstrated either the highest power or at least matched the highest power amongst ATLAS single cutoff thresholds. Larger simulated effect sizes increased power across all thresholds but did not change the relative power exhibited amongst thresholds. Empirical power when using weighted average imputation was similar in comparison to power using best match imputation, although differences in power based on imputation strategy were observed in certain settings when using embeddingMatch (Supplementary Figure 1). In comparison, the estimators proposed by Han *et al* did not capture signal as effectively as ATLAS in all simulation settings. Notably, when records had on average 4 codes, Han *et al*'s estimators failed to capture signal most likely due to noisy linkage probabilities generated from PRL.

We then evaluated ATLAS performance after creating false matches between databases to simulate linkage errors. In doing this, we sought to mimic real-world scenarios where linkage algorithms cannot discern between many pairs of patient records due to a high degree of overlapping features. ATLAS successfully controlled for Type I error in this setting (Supplementary Figure 2). In Figure 4, we show empirical power of ATLAS in presence of false matches. Most notably, use of the weighted average imputation yielded significantly higher power compared to best match imputed variables, and the ATLAS threshold combination test again demonstrated good power relative to its power from single cutoff thresholds. Han *et al*'s M2 estimator yielded comparable power in the context of false matches when using `ludic`, but was generally outperformed by the ATLAS threshold combination test.

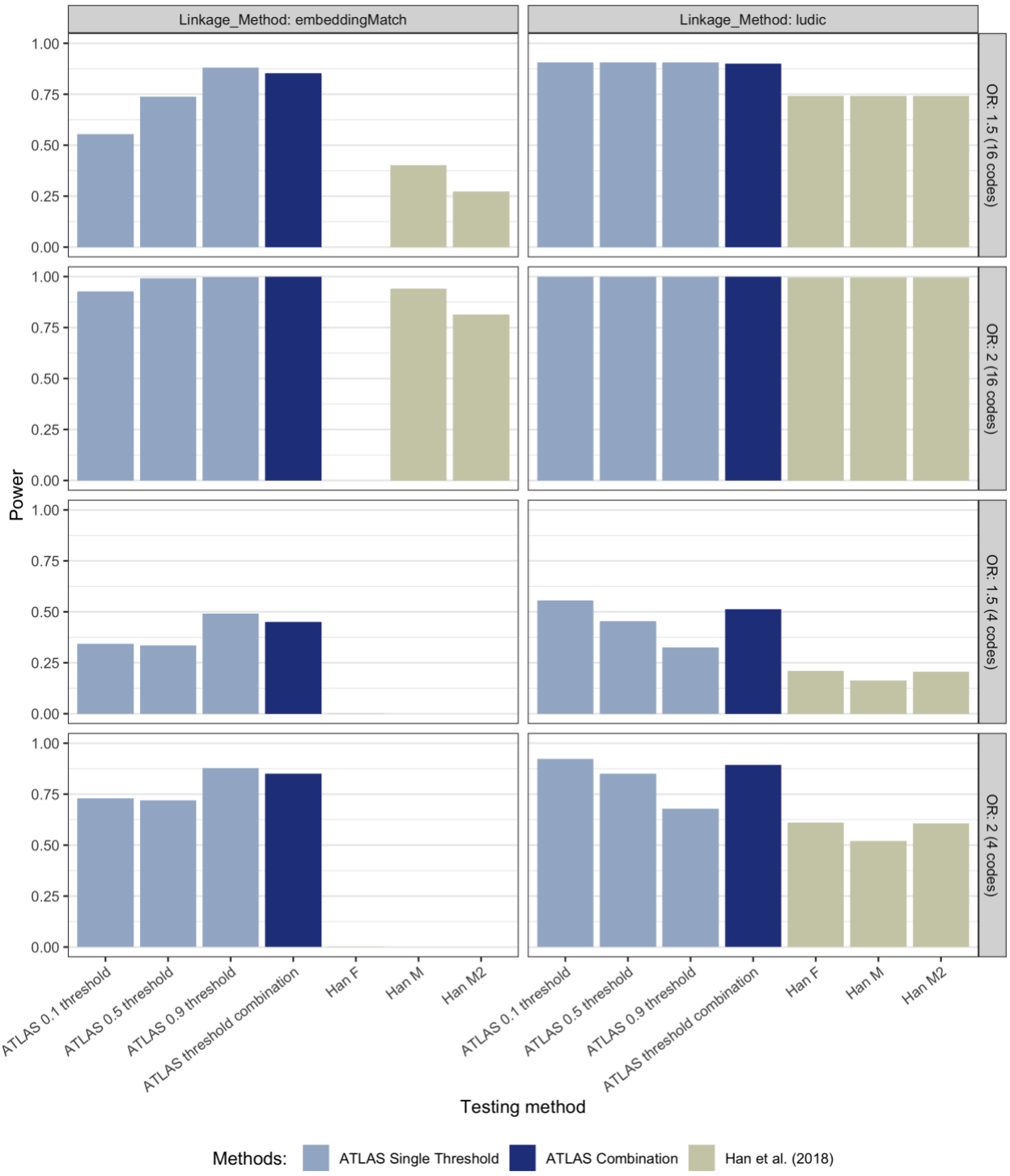

Figure 4: Comparison of empirical power of ATLAS and Han et al. estimators in simulation settings with different effect sizes and average codes per patient record (simulations under H_1).

Results were generated using the best match imputation method. ATLAS power reported for several single cutoff thresholds and the ATLAS threshold combination test.

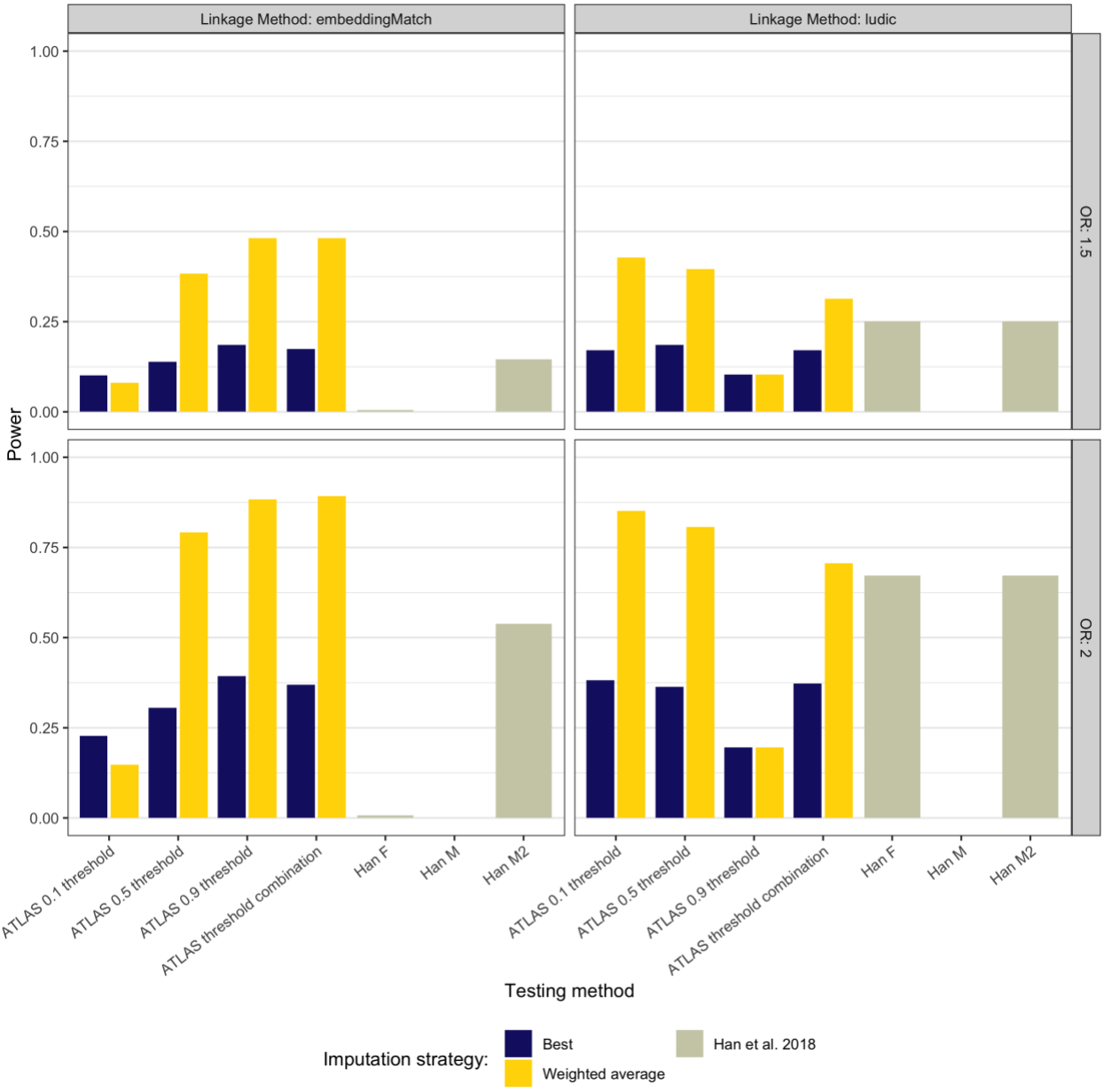

Figure 5: Comparison of empirical power of ATLAS and Han et al. estimators in presence of false matches between databases (simulations under H_1). Simulated databases report on average 16 codes per patient record.

3.3. Real-world genetics study: association of clinical outcomes with rheumatoid arthritis risk alleles among rheumatoid arthritis patients

We performed PRL on 12,838 patient records from RPDR and 1,284 patient records from Crimson. At a conservative threshold of $\rho = 0.9$, we identified 1,157 matching patient records between RPDR and Crimson. Out of 1,157 patients, we further identified 213 patients who have been genotyped in both RPDR and Crimson. Spearman's correlation coefficient between the RPDR and Crimson GRS' for these overlapping patients was estimated to be $\rho = 0.83$ (95% CI: 0.77-0.87, $P = 2 \times 10^{-16}$), suggesting high concordance of genetic data and reliable linkage quality. Univariate association study results using both ATLAS and Han *et al.* estimators are reported in Supplementary Table 1 and 2. As a positive control, we replicated the known association of anti-citrullinated protein antibody (ACPAs) levels with RA GRS using ATLAS and Han *et al*'s estimators [31,32]. In general, ATLAS detected larger effect sizes and smaller p-values compared to any of Han *et al*'s estimators, supporting simulation results that ATLAS is more powerful at detecting associations. For example, for log-transformed rheumatoid factor levels, ATLAS estimated $\beta_{\text{GRS}} = 0.15$ with $p = 0.00$ while Han *et al*'s M2 estimator estimated $\beta_{\text{GRS}} = 0.14$ with $p = 0.09$.

Multivariate association study results using the MGB Biobank and the Crimson linked data study are presented in Table 1 for biomarkers and Table 2 for phenotypes. Effect sizes estimated from both studies were generally concordant. Figure 5 visualizes the difference in adjusted p-values between using only the MGB Biobank cohort for which gold standard mappings were already available and after incorporating additional RA patients with genotype data through the Crimson linked cohort. We demonstrated improved power to detect associations when incorporating

linked data as meta-analysis yielded two additional significant associations (namely log-transformed Erythrocyte sedimentation rate and C-Reactive Protein level). Further, the average unadjusted $-\log_{10}(P\ value)$ among statistically significant outcomes was 6.45 in the MGB Biobank study and 9.65 after meta-analysis with the Crimson linked data study, again demonstrating the potential to increase power when incorporating additional data by linking databases.

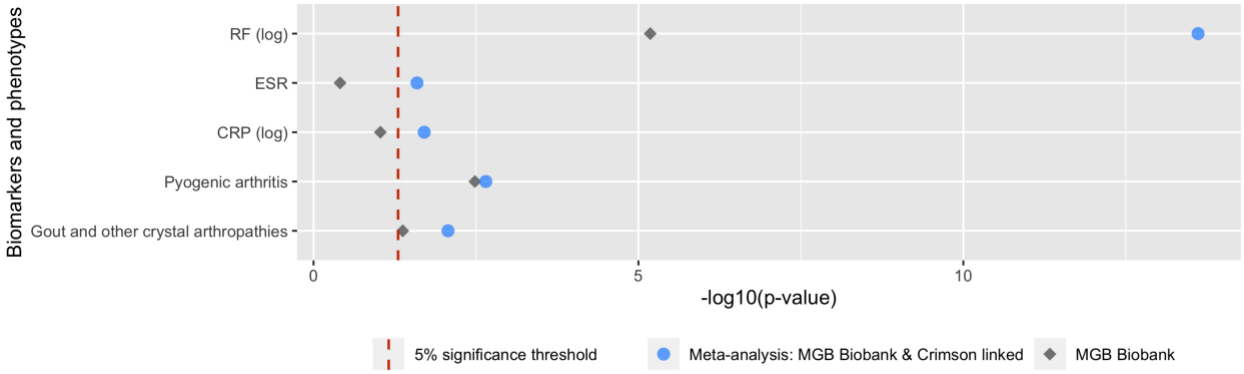

Figure 6: Logarithm transformed P-values from genetic association study using only RA patients with previously available genotype data at MGB Biobank and after incorporating additional RA patients with genotype data through the Crimson linked cohort.

	Beta		**P-value**		
Biomarker	MGB Biobank (SE)	Crimson linked—ATLAS 0.9 threshold (SE)	MGB Biobank	Crimson linked—ATLAS combination	Two study P-value
Anti-citrullinated protein antibodies (log)	0.39 (0.04)	0.31 (0.04)	9.39e-30	0.00e+00	8.54e-38
Rheumatoid factor (log)	0.12 (0.03)	0.16 (0.03)	2.20e-06	0.00e+00	8.14e-15
Erythrocyte sedimentation rate	0.49 (0.58)	1.27 (0.55)	3.91e-01	1.00e-02	2.56e-02
C-reactive protein (log)	0.04 (0.02)	0.04 (0.02)	7.80e-02	3.00e-02	1.65e-02

Table 1: Multivariate association study results of RA GRS and patient biomarkers for the MGB Biobank RA cohort and the Crimson linked RA cohort. Crimson linked RA cohort effect sizes estimated a stringent imputation threshold of 0.9 and P-values estimated using the ATLAS combination test.

	OR		**P-value**		
Phenotype	MGB Biobank (SE)	Crimson linked—ATLAS 0.9 threshold (SE)	MGB Biobank	Crimson linked—ATLAS combination	Two study P-value
Pyogenic arthritis	1.41 (0.11)	1.40 (0.18)	1.64e-03	9.57e-02	1.12e-03
Gout and other crystal arthropathies	0.85 (0.07)	0.85 (0.08)	2.82e-02	2.43e-02	5.68e-03

Table 2: Multivariate association study results of RA GRS and binary phenotypes for the MGB Biobank RA cohort and the Crimson linked RA cohort. Crimson linked RA cohort effect sizes estimated a stringent imputation threshold of 0.9 and P-values estimated using the ATLAS combination test.

4. Discussion

The tremendous amount of biomedical data becoming available for research has led to great interest and demand for linkage of databases, which allows researchers to capture more complete pictures of patient health and conduct novel research studies. Inference using linked data must acknowledge and mitigate bias in estimated effect sizes that are created by linkage errors while retaining good statistical power. To this end, we propose ATLAS as a supervised, robust, flexible, and scalable method that tests for association between variables originally belonging to separate databases. We demonstrate that ATLAS is a valid method that effectively controls for type I error regardless of linkage or imputation method used, and that ATLAS is more powerful than competing methods in a range of linkage and inference settings.

We demonstrate in simulation studies that weighted average imputation of missing variables not only protects against type I error in downstream inference but also preserves good statistical power to detect associations in the presence of linkage errors. Thus, it is the preferable imputation method to use in future studies to mitigate linkaged error induced bias. However, users should be aware that weighted average imputation with lower thresholds using linkage algorithms like `embeddingMatch` may yield lower power as seen in Supplementary Figure 1 when patient records contain few matching features.

During missing variable imputation, instead of selecting ad hoc thresholds above which patients are considered a match or naively imputing variables from pairs of patient records with the highest linkage probability like other proposed methods, ATLAS optimally combines p-values originating from various thresholds. Our results suggest three major advantages to the ATLAS threshold combination test: (1) avoids arbitrarily choosing a threshold for the linkage probability,

thus completely automating the record linkage process; (2) reduces estimation bias by combining p-values estimated from data imputed at different thresholds; and (3) preserves good statistical power regardless of which threshold performs best in a given setting. Both simulation and application study results further demonstrate that the ATLAS threshold combination test substantially outperforms competing methods to detect associations in various settings. When selecting thresholds which are considered in the ATLAS threshold combination test, we caution against using every possible threshold in order to preserve good statistical power and recommend selecting no more than 10 thresholds.

These attractive features of ATLAS enable robust performance in real-world settings, and we demonstrated its utility in our real-world genetic study testing association between RA GRS and clinical outcomes. By meta-analyzing results obtained from MGB Biobank —for which genotype data was already available—and the newly linked Crimson cohort, we were able to increase power and detect two more associations than when using the MGB Biobank cohort alone. Our results suggest that higher GRS are associated with higher levels of inflammatory laboratory biomarkers, increased risk for pyogenic arthritis, and decreased risk for gout, This study demonstrates the potential to detect novel associations by expanding our sample size of patients at no additional cost and without further data collection efforts.

5. Conclusion

In this article, we introduce ATLAS, an automated and flexible algorithm that conducts robust inference using probabilistically linked databases. The ATLAS threshold combination test exhibits high power to detect associations in a range of simulation settings while controlling for type I error, and it exhibits substantial improvement in power and flexibility over existing

inference methods for linked databases. Thus, ATLAS promises to enable novel and powerful research studies using linked data.

6. Funding

This work was supported in part by the U.S. National Institutes of Health Grant U54-HG007963.

7. Author Contributions

All authors made substantial contributions to: conception and design; acquisition, analysis, and interpretation of data; drafting the article or revising it critically for important intellectual content; and final approval of the version to be published.

8. Supplementary Material

Supplementary material is available at Journal of the American Medical Informatics Association online.

9. Conflict of interest statement

None declared.

References

1 Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: Informatics for integrating biology and the bedside. *Journal of the American Medical Informatics Association* 2012;**19**:181–5.

2 Butte AJ. Translational bioinformatics: Coming of age. *Journal of the American Medical Informatics Association* 2008;**15**:709–14.

3 Lesueur F, Azencott C, Laurent M *et al.* A new hybrid record linkage process to render epidemiological databases interoperable: Application to the gemo and genepso studies involving brca1 and brca2 mutation carriers. 2020.

4 Gutman R, Afendulis CC, Zaslavsky AM. A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs. *Journal of the American Statistical Association* 2013;**108**:34–47. doi:10.1080/01621459.2012.726889

5 Neter J, Maynes ES, Ramanathan R. The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association* 1965;**60**:1005–27.

6 Rentsch CT, Harron K, Urassa M *et al.* Impact of linkage quality on inferences drawn from analyses using data with high rates of linkage errors in rural tanzania. *BMC medical research methodology* 2018;**18**:165.

7 Moore CL, Amin J, Gidding HF *et al.* A new method for assessing how sensitivity and specificity of linkage studies affects estimation. *PloS one* 2014;**9**:e103690.

8 Harron K, Goldstein H, Wade A *et al.* Linkage, evaluation and analysis of national electronic healthcare data: Application to providing enhanced blood-stream infection surveillance in paediatric intensive care. *PloS one* 2013;**8**:e85278.

9 Schmidlin K, Clough-Gorr KM, Spoerri A *et al.* Impact of unlinked deaths and coding changes on mortality trends in the swiss national cohort. *BMC medical informatics and decision making* 2013;**13**:1–11.

10 Doidge JC, Harron KL. Reflections on modern methods: Linkage error bias. *International Journal of Epidemiology* 2019;**48**:2050–60.

11 Hof MHP, Zwinderman AH. Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Statistics in Medicine* 2012;**31**:4231–42. doi:10.1002/sim.5498

12 Chipperfield J. A weighting approach to making inference with probabilistically linked data. *Statistica Neerlandica* 2019;**73**:333–50.

13 Dalzell NM, Reiter JP. Regression modeling and file matching using possibly erroneous matching variables. *Journal of Computational and Graphical Statistics* 2018;**27**:728–38.

14 Han Y, Lahiri P. Statistical analysis with linked data. *International Statistical Review* 2019;**87**:S139–57.

15 Hejblum BP, Weber GM, Liao KP *et al.* Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. *Scientific data* 2019;**6**:180298.

16 Bonomi L, Xiong L, Chen R *et al.* Frequent grams based embedding for privacy preserving record linkage. In: *Proceedings of the 21st acm international conference on information and knowledge management*. 2012. 1597–601.

17 Adly N. Efficient record linkage using a double embedding scheme. In: *DMIN*. 2009. 274–81.

18 Shi X, Li X, Cai T. Spherical regression under mismatch corruption with application to automated knowledge translation. *Journal of the American Statistical Association* 2020;1–12.

19 Boutin N, Holzbach A, Mahanta L *et al.* The information technology infrastructure for the translational genomics core and the partners biobank at partners personalized medicine. *Journal of personalized medicine* 2016;**6**:6.

20 Kurreeman F, Liao K, Chibnik L *et al.* Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *The American Journal of Human Genetics* 2011;**88**:57–69.

21 Nalichowski R, Keogh D, Chueh HC *et al.* Calculating the benefits of a research patient data repository. In: *AMIA annual symposium proceedings*. American Medical Informatics Association 2006. 1044.

22 Liao KP, Cai T, Gainer V *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research* 2010;**62**:1120–7.

23 Huang S, Huang J, Cai T *et al.* Impact of icd10 and secular changes on electronic medical record rheumatoid arthritis algorithms. *Rheumatology* 2020.

24 Karlson EW, Boutin NT, Hoffnagle AG *et al.* Building the partners healthcare biobank at partners personalized medicine: Informed consent, return of research results, recruitment lessons and operational considerations. *Journal of personalized medicine* 2016;**6**:2.

25 Gainer VS, Cagan A, Castro VM *et al.* The biobank portal for partners personalized medicine: A query tool for working with consented biobank samples, genotypes, and phenotypes using i2b2. *Journal of personalized medicine* 2016;**6**:11.

26 Okada Y, Wu D, Trynka G *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014;**506**:376–81.

27 Raychaudhuri S, Sandor C, Stahl EA *et al.* Five amino acids in three hla proteins explain most of the association between mhc and seropositive rheumatoid arthritis. *Nature genetics* 2012;**44**:291–6.

28 Denny JC, Ritchie MD, Basford MA *et al.* PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* 2010;**26**:1205–10.

29 Wei W-Q, Bastarache LA, Carroll RJ *et al.* Evaluating phecodes, clinical classification software, and icd-9-cm codes for phenome-wide association studies in the electronic health record. *PloS one* 2017;**12**:e0175508.

30 Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 1995;**57**:289–300.

31 Aggarwal R, Liao K, Nair R *et al.* Anti-citrullinated peptide antibody (acpa) assays and their role in the diagnosis of rheumatoid arthritis. *Arthritis and rheumatism* 2009;**61**:1472.

32 Liao KP, Kurreeman F, Li G *et al.* Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non–rheumatoid arthritis controls. *Arthritis & Rheumatism* 2013;**65**:571–81.