Original Paper

Estimating the Charlson Comorbidity Index From Privacy-Truncated Diagnosis Codes: Single-Center Retrospective Cohort Study Evaluating Five Computational Strategies in 507,949 Inpatient Encounters

Gaetan Kamdje Wabo[1], MSc; Piotr Pawel Sokolowski[1], MD; Mahboubeh Jannesari Ladani[1], MSc; Nandhini Santhanam[1], MSc; Michael Hagmann[1,2], PhD; Thomas Ganslandt[3,4], MD; Fabian Siegel[1], MD

[1]Department of Biomedical Informatics at the Mannheim Institute for Intelligent Systems in Medicine (MIISM), Medical Faculty of Mannheim, University of Heidelberg, Mannheim, Baden-Württemberg, Germany

[2]Department of Medical Statistics, Biomathematics and Information Processing, Medical Faculty of Mannheim, University of Heidelberg, Mannheim, Baden-Württemberg, Germany

[3]Institute of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Bavaria, Germany

[4]Medical Center for Information and Communication Technology, Erlangen University Hospital, Erlangen, Bavaria, Germany

Corresponding Author:

Gaetan Kamdje Wabo, MSc
Department of Biomedical Informatics at the Mannheim Institute for Intelligent Systems in Medicine (MIISM)
Medical Faculty of Mannheim, University of Heidelberg
Theodor-Kutzer-Ufer 1-3, Building 3, Level 4
Mannheim, Baden-Württemberg 68167
Germany
Phone: 49 621 383 8088
Email: gaetankamdje.wabo@medma.uni-heidelberg.de

Abstract

Background: The Charlson Comorbidity Index (CCI) is a standard instrument for risk adjustment in the secondary use of clinical data. Its coding algorithms require *International Classification of Diseases, Tenth Revision*(*ICD-10)* codes with at least 4 characters. Hospital data integration centers may truncate codes to 3 characters for privacy. Truncation makes subcodes ambiguous across Charlson groups and can render the CCI noncomputable.

Objective: The primary objective was to develop, openly release, and evaluate computational strategies that estimate the CCI from 3-character *International Statistical Classification of Diseases, Tenth Revision, German Modification (ICD-10-GM)* codes using national 4-character subcode frequencies from the German Federal Statistical Office (Destatis). The secondary objective was to identify and quantify potential error caused by truncation and error induced by the national prior.

Methods: This single-center retrospective cohort study analyzed 507,949 inpatient encounters of 243,677 patients from University Hospital Mannheim (UMM; 2010-2024). The open-source R (The R Foundation) package *miCCI* (version 1.0.0; developed by Gaetan Kamdje Wabo et al.) implements a deterministic interval estimator (S1), an exact expected-value estimator (S2), multiple imputation (S3), a Dirichlet-smoothed predictive median (S4), and a cross-validated Meta Learner. Age-conditioned priors came from Destatis table 23131-01. The reference standard was the Quan-weighted CCI computed from full-length codes. We report mean absolute error (MAE), root mean square error (RMSE), Bland-Altman limits, Kolmogorov-Smirnov (KS) distance, and exact agreement with patient-level cluster bootstrap 95% CIs. Robustness analyses covered hyperparameters, a local oracle prior, prefixes without Destatis subcodes, Bonferroni-corrected subgroups, and nested cross-validation.

Results: S2, S3, and S4 achieved nearly identical MAE. S2 reached 0.1394 (95% CI 0.1379-0.1408), S3 reached 0.1395 (95% CI 0.1380-0.1408), and S4 reached 0.1390 (95% CI 0.1374-0.1405). All 3 were more accurate than S1 (MAE 0.1678, 95% CI 0.1660-0.1695) and the Meta Learner (MAE 0.1451, 95% CI 0.1437-0.1465). S2 computes exactly the expected score that S3 approximates by sampling, which explains their near-identical errors. S4 preserved the reference distribution best (KS distance 0.023, exact agreement 87.0%). The Meta Learner attained the lowest squared error (RMSE 0.318), also under nested cross-validation. In paired testing within a random subsample of 10,000 patients, S4 at its default setting did not differ from the

best S3 setting. Under an in-sample oracle prior, 73% to 90% of the MAE of S2 to S4 remained, and prior mismatch accounted for the rest. After Bonferroni correction, 10 of 20 *ICD-10* chapters retained a decisive best strategy.

Conclusions: In this single-center cohort, frequency-informed estimation reconstructed the CCI from 3-character codes with an MAE of about 0.14 points (S2: 0.1394, 95% CI 0.1379-0.1408; S3: 0.1395, 95% CI 0.1380-0.1408; S4: 0.1390, 95% CI 0.1374-0.1405). Expected-value estimation (S2) is the most parsimonious choice for encounter-level accuracy. The integer-valued median (S4) can suit analyses that depend on the score distribution. Use in clinical research remains a potential application that requires multisite validation.

JMIR Med Inform 2026;14:e107041; doi: 10.2196/107041

Keywords: Charlson Comorbidity Index; *ICD-10-GM*; data truncation; privacy-preserving data sharing; multiple imputation; Bayesian inference; comorbidity scoring; administrative health data; data quality; secondary data use; *International Statistical Classification of Diseases, Tenth Revision, German Modification*; digital health; medical computer science; *miCCI* package; health services research

Introduction

Background

Comorbid conditions shape clinical trajectories. Patients with several coexisting diseases face higher mortality, longer stays, and greater resource use [1,2]. Quantifying this burden reproducibly is essential for risk adjustment in epidemiology, health services research, and quality benchmarking. A recent systematic review identified 22 multimorbidity measures based on *International Statistical Classification of Diseases* (*ICD)* codes in administrative data and found marked heterogeneity in their construction and validation [3]. Within this landscape, the Charlson Comorbidity Index (CCI) remains the predominant instrument [4,5]. National norms of the CCI have now been reported for hospitalized adults [6]. The CCI also informs clinical classification, for example as a risk threshold after spinal metastasis surgery [7] and as a prognostic factor in spine tumor and hip-fracture patients [8,9].

Automated scoring became possible once the CCI was mapped to administrative coding systems. Deyo et al [10] created the first *International Statistical Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)* mapping. Romano et al [11] refined it. Quan et al [12] extended it to the *International Statistical Classification of Diseases*, *Tenth Revision* (*ICD-10)*, and Sundararajan et al [13] validated an independent *ICD-10* translation. A country-specific mapping for German administrative data has recently been published [14].

These mappings presuppose codes with 4 or 5 characters. The prefix E11 (type 2 diabetes mellitus) is ambiguous because E11.4 maps to complicated diabetes (weight 2), whereas E11.9 maps to uncomplicated diabetes (weight 1). Similar ambiguities arise for liver disease, myocardial infarction, and malignancies [12]. Truncation to 3 characters is a generalization step that data integration centers may apply before releasing data to researchers or federated networks [15-17]. The analytical cost of such privacy transformations on clinically meaningful structures remains insufficiently characterized [18,19]. We call this operation truncation and do not equate it with formal anonymization, because residual risk depends on all released attributes.

Prior work has not closed this gap. The Stausberg Index [20] operates on 3-character *International Statistical Classification of Diseases, Tenth Revision, German Modification* (*ICD-10-GM)* codes but defines a new instrument rather than reconstructing the CCI. The R packages *comorbidity* [21] and *icd* [22] require full-length codes. To our knowledge, no open tool estimates the CCI from truncated codes using population frequencies as priors.

The required priors are public. Destatis publishes inpatient case counts for every 4-character *ICD-10-GM* code across 22 age bands in table 23131-01 [23]. These counts define the probability of each subcode within its prefix. Because the prevalence of individual Charlson conditions varies with age [6], age-specific counts can sharpen that prior.

This study introduces the open-source R package *miCCI* [24]. The contribution is not a new score. It is a method that keeps a standardized index computable after privacy-driven data minimization.

Objectives and Research Questions

The primary objective was to develop, release, and evaluate strategies that estimate the CCI from *three*-character *ICD-10-GM* codes. The secondary objective was to decompose the reconstruction error into its sources. The exploratory objective was to locate subgroups in which strategies differ. Four research questions (RQs) followed from these objectives.

- RQ1: How accurately can the encounter-level CCI be reconstructed from 3-character codes, and how do 5 strategies compare?
- RQ2: Do the strategies preserve the marginal distribution of the reference standard CCI?
- RQ3: How much reconstruction error stems from truncation itself, from the national prior, from subcodes missing in the prior, and from Monte Carlo settings?
- RQ4 (exploratory): Which strategy performs best within *ICD-10* chapters, active Charlson groups, and multimorbidity patterns once multiplicity is controlled?

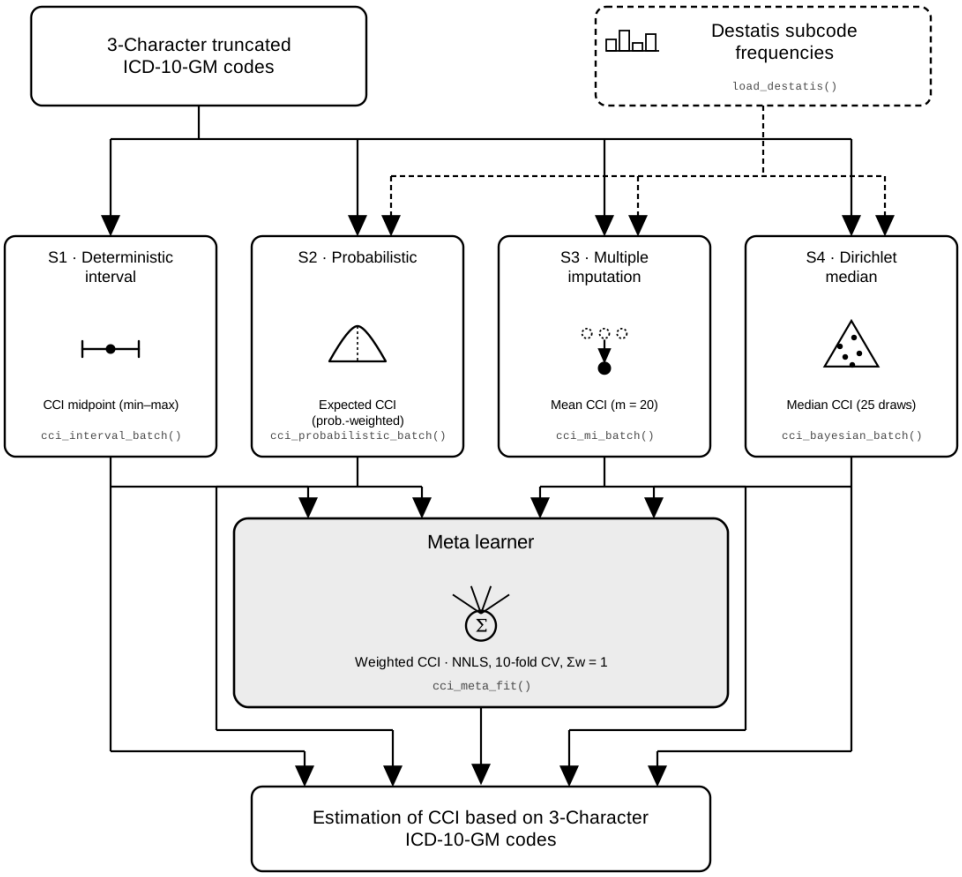
Methods

Study Design and Cohort

This single-center retrospective cohort study used inpatient encounters from University Hospital Mannheim (UMM) admitted between January 2010 and September 2024. Figure 1 summarizes the workflow. The encounter was the unit of analysis because a truncated code list is generated once per discharge episode. Pseudonymized patient identifiers were used for statistical inference only (see Statistical Evaluation).

The extract holds 1 row per encounter. The encounter number (falnr) is the key, and all diagnoses of an encounter are stored in one delimited field. Deduplication on falnr therefore removes only repeated rows of the same encounter. Readmissions carry their own encounter numbers and are retained. The final cohort of 507,949 encounters from 243,677 patients (2.08 encounters per patient) is consistent with this schema. Rows were excluded stepwise if the admission date was missing or unparseable, if no diagnosis code was recorded, if length of stay or age was missing or negative, or if admission fell outside the study period. Duplicate rows of an encounter number were removed last. The number of rows excluded at each step is reported in tab_cohort_00_selection_flow.csv (Multimedia Appendix 1). Sex was not extracted, in line with the data-minimization principle of the institutional data integration center.

Figure 1. Workflow of the *miCCI* reconstruction. 3-character truncated *International Statistical Classification of Diseases, 10th Revision, German Modification* (*ICD-10-GM*) codes and Destatis subcode frequencies (dashed arrows) feed 4 base strategies. S1 (strategy 1 [deterministic interval]) uses only the set of subcodes under each prefix. S2 (strategy 2 [exact expected value]) to S4 (strategy 4 [Dirichlet-smoothed predictive median]) also use their frequencies. The Meta Learner combines S1 to S4 with nonnegative weights that sum to 1. Each path yields a CCI estimate from the same 3-character codes. The labels *M*=20 and 25 draws denote the default settings. CCI: Charlson Comorbidity Index; CV: cross-validation; NNLS: nonnegative least squares; S1: strategy 1 (deterministic interval); S2: strategy 2 (exact expected value); S3: strategy 3 (multiple imputation); S4: strategy 4 (Dirichlet-smoothed predictive median).



Inclusion of Pediatric Encounters

The cohort contains 9977 encounters of patients younger than 18 years (1.96%). The CCI was developed and validated in adults, and we make no claim about its prognostic validity in children. Adult indices discriminate poorly in pediatric populations. Sun et al [25] reported C statistics of 0.522 to 0.640 for adult scores in children, compared with 0.718 for a dedicated Pediatric Index. Ikeda Kurakawa et al [26] found a C statistic of 0.544 for the CCI and 0.720 for a pediatric Children Comorbidity Score. Childhood multimorbidity itself remains underexplored and lacks standardized measurement [27].

We nevertheless retained these encounters for 3 reasons. First, the target quantity of this study is fidelity of score reconstruction against an algorithmic reference, which is defined for every encounter regardless of age. Second, data integration centers truncate pediatric records through the

same pipelines, so excluding them would leave an unquantified error in released data. Third, Destatis reports 5 pediatric age bands labeled <1, 1-5, 5-10, 10-15, and 15-18 years. Each label runs from the lower bound to under the upper bound, so the bands cover <1, 1-4, 5-9, 10-14, and 15-17 completed years. The age-conditioned prior thus applies without extrapolation. Pediatric reconstruction is developed further under Future Directions.

Ethical Considerations

The study was approved by the Medical Ethics Commission II, Medical Faculty Mannheim, Heidelberg University (approval ID: 2024-885). All data were pseudonymized at extraction. The study followed the Declaration of Helsinki and German data protection law. Participants received no compensation.

Definitions

Truncation

Truncation replaces a 4- or 5-character code with its 3-character parent. It is a hierarchical generalization step used under privacy models such as k-anonymity [15,17].

Reference Standard CCI

The reference is the nonage-adjusted, 17-condition Quan-weighted CCI [12] computed from the same center's full-length codes. Three hierarchical rules suppress the milder form when both forms are present (mild and severe liver disease, uncomplicated and complicated diabetes, and nonmetastatic and metastatic malignancy). The *ICD-10-GM* mapping follows the German adaptation by Sokołowski et al [14] and is distributed as a machine-readable JSON file (codes_quan_orig.json; Multimedia Appendix 2). Reference scoring is implemented in miCCI/R/01_cci_mapping.R (Multimedia Appendix 2). The reference is algorithmic, not an adjudicated clinical diagnosis, and is therefore termed reference standard rather than gold standard.

Reconstruction

Reconstruction means score-level estimation of the CCI that the full-length codes would have produced. It does not denote recovery of the individual subcodes.

Population Frequencies

The prior is the 2023 release of Destatis table 23131-01 [23], which reports hospital diagnoses by 4-character code and 22 age bands. Any table with the same schema can replace it through the freq_table argument.

Silent Gap and Degenerate Prefixes

The 2024 *ICD-10-GM* catalog [28] contains 9306 four-character codes (icd_catalogue.csv, Multimedia Appendix 2). A catalog subcode absent from Destatis lies in the silent gap and receives zero probability. A prefix without any child in Destatis is termed degenerate. This occurs when all its subcodes lie in the silent gap or when the prefix is itself a terminal 3-character code, such as C61 (prostate cancer) or F03 (unspecified dementia).

Prior Construction

A single precomputation (miCCI/R/02_destatis.R, Multimedia Appendix 2) runs once per Destatis release. For each prefix, it stores the probability of each child, computed within each of the 22 age bands and marginally, and the Charlson group of each child. Each encounter age is mapped to its Destatis band. When that band carries no mass for a prefix, the marginal probabilities are used. We chose the age-conditioned prior because the prevalence of individual Charlson conditions varies with age [6]. The package documentation (README.md, Multimedia Appendix 2) illustrates the effect for an isolated E11 code. Its S2 expected score rises from 1.23 at age 35 to 1.65 at age 78, because complicated diabetes becomes more frequent with age. Table 23131-01 belongs to the Destatis hospital diagnosis statistics (Hospital Patient Diagnosis Statistics, registered under statistical code 23131 in the Unified Register of Federal and State Statistics, EVAS 23131), which records only the principal diagnosis of each discharged inpatient. The prior therefore reflects subcode frequencies among admissions for which the condition was the principal reason for admission. Charlson conditions, by contrast, are predominantly coded as secondary diagnoses. Codes that German coding rules restrict to secondary positions, such as asterisk codes, are thus largely absent from the table. We treat this difference in diagnosis role as part of prior mismatch.

Code multiplicity was preserved. Each coded diagnosis position keeps its own prefix slot, so E11 coded twice yields 2 independent completions. Two positions under one prefix can represent 2 different subcodes, for example I25.1 and I25.2. Collapsing them to one slot would undercount comorbidities that share a prefix.

Prefixes follow 3 explicit rules. A resolved prefix draws from its Destatis children, and silent-gap subcodes receive zero probability. A degenerate prefix stands for itself with probability 1 and is matched against the Quan patterns in both directions. Any Charlson group with a pattern under that prefix is therefore activated. An empty entry contributes nothing. The degenerate rule is neither an error nor a zero. It prevents silent loss of valid three-character diagnoses. It also scores the AIDS codes B20 to B24, which have no children in Destatis, identically under every strategy.

Computational Reconstruction Strategies

Table 1 summarizes the 5 strategies. Formal definitions follow.

The table states design properties. Empirical performance is reported in the Results section. All strategies apply the Quan exclusion rules and the degenerate-prefix rule. Implementation files are named in the corresponding Methods subsections.

Table 1. Overview of the 5 reconstruction strategies implemented in *miCCI*.

Strategy	Principle	Output support	Inputs beyond the truncated codes	Expected advantages	Potential limitations
S1[a] interval	Midpoint of conservative lower and upper CCI[b] bounds for each encounter	Half-integer	Subcode sets of each prefix	Explicit lower and upper bounds. No frequency table needed	Ignores subcode frequencies. Upper bound conservative by construction
S2[c] expected value	Exact expected CCI under the prior, with hierarchical exclusion handled in closed form	Continuous	Subcode frequencies (prior)	Deterministic and fast. Exposes per-group probabilities	Assumes independent completion of coded positions. Rarely returns an integer
S3[d] multiple imputation	Mean CCI over m random completions drawn from the prior	Continuous	Subcode frequencies (prior)	Simulates plausible complete coding. Extensible to any score	Monte Carlo approximation of the S2 expectation. Adds sampling noise
S4[e] Dirichlet median	Median CCI over n completions with Dirichlet-perturbed probabilities	Integer	Subcode frequencies (prior)	Same integer support as the reference score	Sensitive to prior strength and aggregator. Sampling noise
Meta learner	Nonnegative least-squares combination of S1 to S4	Continuous	Calibration set with full-length codes	Can exploit complementary errors of the base strategies	Requires site-specific refitting. Optimized for squared error

[a]S1: strategy 1 (deterministic interval).
[b]CCI: Charlson Comorbidity Index.
[c]S2: strategy 2 (exact expected value).
[d]S3: strategy 3 (multiple imputation).
[e]S4: strategy 4 (Dirichlet-smoothed predictive median).

Strategy 1 (S1): Deterministic Interval

For each prefix, S1 classifies each Charlson group as certain (every child maps to it), possible (some children map to it), or absent. The lower bound sums the weights of certain groups. The upper bound sums the weights of certain or possible groups. Hierarchical exclusion is applied to both bounds. In all 3 Quan pairs, the parent outweighs the child, so suppression removes the infeasible sum within each pair. Nonhierarchical alternatives are deliberately not resolved. Two groups reachable only through different children of one coded position, such as heart failure (I13.0) and kidney disease (I13.1), can both enter the upper bound.

$$CCI_{S1} = (CCI_{min} + CCI_{max})/2 \quad (1)$$

CCI_min and CCI_max are the suppressed bounds of a conservative envelope. They contain the reference score if the original subcodes are listed in the prior. S1 uses the children of each prefix but not their frequencies (miCCI/R/10_cci_interval.R, Multimedia Appendix 2).

Strategy 2 (S2): Exact Expected Value

For prefix p, age band a, and group g, the per-position hit probability is $q_{pg} = \sum_{c\pi}(c \mid p, a) \cdot 1[c \in g]$, where π is the prior and the sum runs over the children c of p. Coded positions are completed independently. A group without a severity parent is then active with probability

$$P(g) = 1 - \Pi_p(1 - q_{pg})_p^k \quad (2)$$

where k_p is the number of positions coded with prefix p. For a child group c with parent set Par(c), let q_par be the probability that a completion maps to any parent and q_uni the probability that it maps to c or to any parent. The probability that c scores after suppression is

$$P_{eff(c)} = \Pi_p(1 - q_{par,\,p})_p^k - \Pi_p(1 - q_{uni,\,p})_p^k \quad (3)$$

$$CCI_{S2} = \sum_g w_g \cdot P_{eff(g)} \quad (4)$$

Here w_g is the Quan weight, and P_eff(g) equals P(g) for groups without a parent. Equation 3 requires no independence between branches of the same prefix. It therefore handles their mutual exclusivity exactly. Treating child and parent activation as independent events would overstate the suppression of the child. Equation 4 is the exact expected reference score of a random completion under the prior (miCCI/R/20_cci_probabilistic.R, Multimedia Appendix 2).

Strategy 3 (S3): Multiple Imputation

S3 draws one subcode per coded position from the prior, scores the completed list with the reference algorithm, and repeats this M=20 times.

$$CCI_{S3} = (1/m) \cdot \sum_i CCI_{ref}(D_i) \quad (5)$$

D_i is the i-th completed code list. S3 is therefore a Monte Carlo estimator of the quantity that S2 computes in closed form. We chose M=20 because it is the pilot size in the two-stage rule of von Hippel [29], and its sensitivity was evaluated (miCCI/R/30_cci_mi.R, and analysis/08_hyperparameter_sensitivity.R, Multimedia Appendix 2). Rubin combining rules were not applied because no variance decomposition is reported [30,31].

Strategy 4 (S4): Dirichlet-Smoothed Predictive Median

For each prefix with K children, S4 samples a probability vector from a Dirichlet distribution with concentration α_k

= α_0 · π_k+0.5, where α_0=10 controls prior strength and 0.5 is a Jeffreys-type smoothing term [32]. It then draws one subcode per position and scores the completed list. After n=25 draws, the estimate is the median.

$$CCI_{S4} = median_d CCI_{ref}(D_d) \quad (6)$$

The truncated record supplies no subcode counts, so no likelihood update occurs. The result is a smoothed prior predictive distribution, and Bayesian is used as shorthand. The median of an odd number of integer scores is itself an integer. S4 output is therefore discrete and shares the support of the reference. The median was chosen because it is robust to the added between-draw variability of the Dirichlet layer (miCCI/R/40_cci_bayesian.R, Multimedia Appendix 2).

Meta Learner

The Meta Learner combines S1 to S4 with the Super Learner [33] through the *SuperLearner* R package [34]. Each base strategy enters as a pass-through learner that returns its precomputed column. Ten-fold cross-validation produces out-of-fold predictions, and nonnegative least squares (NNLS) fits weights that are nonnegative and sum to one (miCCI/R/60_cci_meta.R, Multimedia Appendix 2).

$$CCI_{Meta} = w_1 \cdot S1 + w_2 \cdot S2 + w_3 \cdot S3 + w_4 \cdot S4 \quad (7)$$

NNLS minimizes squared error, whereas strategies are ranked mainly by mean absolute error (MAE). We kept the squared-error objective deliberately. Phillips et al [35] recommend choosing the loss according to the intended use of the ensemble. Here the Meta Learner serves as a diagnostic of complementarity between strategies. Squared loss penalizes large errors, which matter most when the CCI enters regression-based risk adjustment. We therefore evaluate the Meta Learner on its own objective root mean square error (RMSE) and on MAE, and we treat an MAE-trained ensemble as future work.

The weights were fitted on the full cohort. The pass-through learners have no parameters, so the only fitted quantities are 4 weights constrained to the simplex, that is, 3 free parameters [33,34]. For a linear fit with d free parameters, the expected optimism of the in-sample squared error is approximately $2d\sigma^2/n$ [36], or 2.5×10^{-6} for d=3, a residual variance of 0.101, and n=243,677 patients. We verified this by patient-level nested cross-validation [36,37]. Patients were split into 10 outer folds. For each fold, the Meta Learner was refit on the other 9 folds, including its 10 inner folds, and applied to the held-out patients. Every encounter was thus predicted by a model that never saw its patient. S1 to S4 fit no parameters on the cohort, so their predictions are already out of sample. Out-of-fold metrics and paired differences used the same patient-level cluster bootstrap described under Statistical Inference. A full-cohort refit reproduced the published weights to within 10^{-5} (analysis/10_meta_outer_cv.R, Multimedia Appendix 2).

Statistical Evaluation

Point Accuracy and Agreement

For RQ1, we computed MAE, RMSE, R^2, and bias (mean of predicted minus reference). Agreement across the score range was assessed with Bland-Altman analysis [38]. For each encounter, the difference d_i = ŷ_i – y_i was plotted against the mean (ŷ_i+y_i)/2. The limits of agreement are bias±1.96 SD of d_i. Because the reference is discrete and errors are heteroscedastic, the limits are descriptive. MAE served as the primary ranking metric because it is expressed in score points and, unlike RMSE, is not dominated by rare large errors [39]. RMSE and R^2 complement it for squared-error behavior. All accuracy and agreement metrics are computed in miCCI/R/95_report.R (Multimedia Appendix 2).

Distributional Fidelity

For RQ2, we report the Kolmogorov-Smirnov (KS) distance between predicted and reference distributions. We omit its *P*-value, which is uninformative at this sample size. We also report the share of encounters within ±0.5, ±1, and ±2 points of the reference. Exact agreement is the share of encounters whose prediction equals the reference value exactly. Value-specific exact agreement is the same share among encounters with reference value v. Exact agreement requires no rounding, and half-integer outputs count as nonmatches. Binned score frequencies are reported only as supplementary information. Their binning rounds halves away from zero (function .round_half_up in miCCI/R/70_qa_distribution.R, Multimedia Appendix 2), because rounding half to even would shift S1 midpoints preferentially onto even bins.

Mass Conservation Ratio

For each Charlson group g, the reference mass is M_g^ref = (w_g / N) · Σ_i 1[g scores in encounter i]. The strategy mass is M_g^s = (w_g / N) · Σ_i p_ig^s, where p_ig^s is the activation of g after suppression. This activation equals 1 or 0.5 for S1, P_eff(g) for S2, and the share of draws in which g scores for S3 and S4. The mass conservation ratio is R_g=M_g^s / M_g^ref, where 1 denotes perfect attribution. Ratios within ±15% of 1 were considered conserved, adopting the ±15% relative accuracy criterion of quantitative method validation [40]. Summed over groups, the strategy masses reproduce the mean score exactly for S2 and S3. They deviate by 0.002 points for S1 and 0.005 points for S4, whose median is not additive. The ratio is computed in miCCI/R/70_qa_distribution.R (Multimedia Appendix 2).

Statistical Inference

Readmissions induce correlation between encounters of the same patient. All CIs therefore come from a cluster bootstrap that resamples patients with replacement and keeps all their encounters together [41,42]. We used B=1000 replicates for global metrics and B=500 for per-stratum metrics. Strategies were compared with a paired bootstrap of the MAE difference, which recomputes both strategies on the same replicate [43]. This absorbs the strong correlation induced by evaluation on identical encounters (miCCI/R/85_bootstrap_ci.R,

Multimedia Appendix 2). Temporal stability was assessed as MAE per calendar year.

Error Source Analyses

Three analyses addressed RQ3. First, a local oracle prior was built from the cohort's own four-character codes, and S1 to S4 were rerun (analysis/07_prior_provenance.R, Multimedia Appendix 2). Following the oracle concept of Donoho and Johnstone [44], this prior uses information unavailable in practice and thus benchmarks an estimator with ideal knowledge. The MAE under this prior estimates the error that remains when the prior matches the cohort's own pooled code distribution. It therefore contains truncation and any heterogeneity that one pooled prior cannot capture. Complete pooling masks variation between subgroups [32], and comorbidity coding in German hospitals has changed over time and differs between hospitals [45]. Coding drift between years thus remains in this residual. The difference from the national-prior MAE estimates the error caused by prior mismatch. The oracle is estimated in sample, and in-sample error is optimistic by construction [36]. Its error is therefore an optimistic benchmark, not a deployable alternative. A group deviation was attributed to prior mismatch, which combines case mix and diagnosis role (see Prior Construction), when its mass ratio lay outside ±15% under the national prior and within ±15% under the local prior. It was attributed to truncation when it remained outside.

Second, the Destatis table was compared with the 2024 catalog (analysis/05_destatis_vs_catalogue.R and analysis/06_silent_gap_analysis.R, Multimedia Appendix 2). We counted degenerate prefixes in the cohort and compared S1 and S2 under 2 rules: silent dropping of these prefixes and self-fallback. This comparison used the marginal prior.

Third, Monte Carlo settings were varied over $m \in \{5, 10, 20, 40, 100\}$ for S3 and a full grid of $n \in \{5, 15, 25, 51, 101\}$, $\alpha_0 \in \{1, 10, 100\}$, and median or mean aggregation for S4 (30 cells). Every S4 cell was compared with the best S3 setting by paired cluster bootstrap, with Bonferroni adjustment over the 30 cells (analysis/08_hyperparameter_sensitivity.R, Multimedia Appendix 2). Choosing the best S3 setting as comparator makes this test conservative for S4.

This grid ran on a random subsample of 10,000 patients with all 20,973 of their encounters. We chose this design for 4 reasons. First, the grid requires up to 100 imputation rounds for S3 and 101 draws for each of 3 prior strengths for S4. On the full cohort, this would multiply the 8.2-hour compute stage several-fold. Second, sampling whole patients keeps the cluster structure intact, so the cluster bootstrap remains valid. Third, each setting was sliced from one run with the largest number of draws per α_0. Settings therefore share random numbers and differ only in the parameter under study, which reduces noise between cells. Fourth, the subsample yielded descriptively comparable default MAEs of 0.1400 for S3 and 0.1378 for S4, versus 0.1395 (95% CI 0.1380-0.1408) and 0.1390 (95% CI 0.1374-0.1405), respectively, in the full cohort.

Exploratory Subgroup Analysis

For RQ4, encounters were stratified by dominant *ICD-10* chapter, by active reference Charlson group, and by the 30 most frequent multimorbidity patterns. The dominant chapter is the chapter most frequent among the distinct three-character prefixes of an encounter. Strata with fewer than 50 encounters were omitted. In each stratum, the strategy with the lowest MAE was compared with the runner-up by paired cluster bootstrap. A stratum was decisive when the CI of the difference excluded zero. To control multiplicity, each family of K strata received Bonferroni-adjusted intervals at level 1-0.05 /K [46]. Both levels were read from the same replicate vector, so a Bonferroni-decisive stratum is always decisive at 95% (miCCI/R/80_stratified_eval.R and analysis/09_multiplicity_correction.R, Multimedia Appendix 2).

Implementation

All strategies were implemented in R version 4.6.0 [47] and released as the MIT-licensed *miCCI* package, version 1.0.0 [24]. Computation relies on data.table 1.18.4, and the ensemble on SuperLearner 2.0-42 with nnls 1.6. The complete session information, configuration, and seed are recorded in run_manifest.json (Multimedia Appendix 3). This file is kept as written by the pipeline, except that the 2 local file paths were masked for data protection reasons. The results were produced with a development build labeled 1.1.0 that is computationally equivalent to release v1.0.0. The script run_full_cohort.R (Multimedia Appendix 2) executes the compute stage, the reporting stage, and analysis scripts 03 to 10, and stops unless invariant checks pass. The full-cohort compute stage ran in 8.2 hours (29,665 seconds) on a Windows 11 × 64 workstation. Checkpoints are keyed to a fingerprint of the cohort and of the configuration, so any setting change invalidates stale intermediate results. The package ships 63 unit test blocks (miCCI/tests/testthat, Multimedia Appendix 2) covering reference scoring, the 3 suppression rules, equivalence of single and batch paths, the S1 bounds and their conservative envelope, the exact S2 expectation, degenerate prefixes, the age-conditioned prior, code multiplicity, the bootstrap functions, and checkpointing. Each Methods subsection above names its implementing file. Each Results subsection below names the output files that reproduce its numbers.

Results

Results follow the research questions, including the accuracy (RQ1), distributional fidelity (RQ2) with its group-level and temporal complements, error sources (RQ3), and exploratory subgroups (RQ4).

Cohort Characteristics

The final cohort comprised 507,949 encounters from 243,677 patients (Tables 2 and 3). The raw extract held 1,097,091 rows. Exclusions removed 201,928 rows without a diagnosis code (18.4%), 685 with missing or negative length of stay, 6 with missing or negative age, and 379,503 with admission outside the study period (34.6%). No

row lacked an admission date. Removing 7,020 duplicate rows (0.6%) yielded the final cohort (tab_cohort_00_selection_flow.csv, Multimedia Appendix 1). These rows repeated 6,282 encounter numbers, none of which carried differing diagnosis strings. The median age was 63 years (IQR 48-74), the median number of diagnoses per encounter was 6 (IQR 4-10), and the median length of stay was 4 days (IQR 2-9). Annual volume fell from 45,506 encounters in 2010 to 24,986 in the 9 months of 2024. The reference CCI had a mean of 1.578 (SD 2.052) and a median of 1 (IQR 0-2). A score of zero applied to 45.31% of encounters (tab_04_qa_score_frequency.csv, Multimedia Appendix 3). Full descriptives are provided in Multimedia Appendix 1.

Source files: tab_cohort_01_overall.csv, tab_cohort_02_age_distribution.csv, and tab_cohort_03_yearly_volume.csv (Multimedia Appendix 1). The patient count is n_clusters in tab_01_strategy_metrics_ci.csv, and the reference CCI is from tab_02_descriptive_statistics.csv (Multimedia Appendix 3). Produced by analysis/03_cohort_descriptive.R (Multimedia Appendix 2).

Table 2. Cohort characteristics (University Hospital Mannheim, 2010-2024).

Variable	Mean (SD)	Median (IQR)
Encounters[a]	—[b]	—
Patients[c]	—	—
Age (years)	59.9 (19.0)	63 (48-74)
ICD-10-GM[d] diagnoses per encounter	8.4 (7.9)	6 (4-10)
Length of stay (days)	9.3 (14.9)	4 (2-9)
Reference standard CCI[e]	1.578 (2.052)	1 (0-2)

[a]N=507,949.
[b]Not available.
[c]N=243,677.
[d]*ICD-10-GM: International Classification of Diseases, Tenth Revision, German Modification*.
[e]CCI: Charlson Comorbidity Index.

Table 3. Cohort characteristics (University Hospital Mannheim, 2010-2024).

Age group (in years)	Encounters and shares, n (%)
<18	9977 (1.96)
18-29	36,451 (7.18)
30-39	36,011 (7.09)
40-49	51,695 (10.18)
50-59	84,216 (16.58)
60-69	104,914 (20.65)
70-79	112,032 (22.06)
80-89	63,478 (12.50)
≥90	9175 (1.81)

Point Accuracy (RQ1)

S2, S3, and S4 reached nearly identical MAE (Tables 2–4). Their 95% CIs overlap almost completely, and the largest difference among them is 0.0005 points. S1 had an MAE of 0.1678 (95% CI 0.1660-0.1695), and the Meta Learner had an MAE of 0.1451 (95% CI 0.1437-0.1465). Relative to S1, S2 to S4 reduced the MAE by about 17%. The Meta Learner reached the lowest RMSE (0.318) and the highest R^2 (0.976). S4 had a higher RMSE (0.400) than S2 (0.341) and S3 (0.345), which reflects its integer output. All strategies overestimated the reference slightly, with bias between 0.040 (S3) and 0.085 (S1). Unless stated otherwise, metrics in this section are from tab_01_strategy_metrics_ci.csv (Multimedia Appendix 3).

S2 and S3 agreed closely on every metric, with MAE 0.1394 (95% CI 0.1379-0.1408) versus 0.1395 (95% CI 0.1380-0.1408) and mean scores of 1.6182 (SD 2.064) versus 1.6181 (SD 2.064). This agreement follows from their construction, because S3 approximates by sampling the expectation that S2 computes exactly. The residual difference is Monte Carlo noise, which also explains the slightly higher RMSE of S3.

The Meta Learner results held out of sample. Under patient-level nested cross-validation, its MAE was 0.1451 (95% CI 0.1437-0.1465), 2.0×10^{-6} points above the in-sample value (95% CI 0.4×10^{-6} to 3.6×10^{-6}). The observed optimism of the mean squared error (MSE) was 3.8×10^{-6} (out-of-sample MSE 0.101105, 95% CI 0.099579-0.102512; in-sample MSE 0.101101, 95% CI 0.099577-0.102507), of the same order as the expected 2.5×10^{-6}. Out of sample, the Meta Learner remained behind S2, S3, and S4 by 0.0058 (95% CI 0.0053-0.0063), 0.0057 (95% CI 0.0051-0.0062), and 0.0062 (95% CI 0.0054-0.0070) points, and each paired 95% CI excluded zero (tab_mo_02_metrics_ci.csv to tab_mo_04_summary.csv, Multimedia Appendix 4).

Reference standard CCI: mean 1.578 (SD 2.052). CIs were estimated using patient-level cluster bootstrapping (B=1000 replicates). Bland-Altman limits are bias ±1.96 SD of the differences. The KS *P* value is omitted because it is uninformative at this sample size. Bland-Altman limits were derived from tab_01_strategy_metrics_ci.csv as bias ±1.96×√(RMSE² – bias²). Source files: tab_01_strategy_metrics_ci.csv, tab_02_descriptive_statistics.csv, tab_05_qa_ks_distance.csv, tab_06_qa_posterior_coverage.csv, and meta_weights.csv (Multimedia Appendix 3), and tab_14_exact_agreement_summary.csv (Multimedia Appendix 5). Produced by miCCI/R/95_report.R (Multimedia Appendix 2).

Table 4. MAE[a] across 5 computational strategies[b].

Computational strategy	MAE (95% CI)
S4 – Dirichlet-smoothed predictive median	0.1390 (0.1374-0.1405)
S2 – Expected-value estimator	0.1394 (0.1379-0.1408)
S3 – Multiple imputation	0.1395 (0.1380-0.1408)
Meta Learner	0.1451 (0.1437-0.1465)
S1 – Deterministic interval estimator	0.1678 (0.1660-0.1695)

[a]MAE: mean absolute error.
[b]Results are based on 507,949 inpatient encounters from 243,677 patients. The 95% CIs were estimated using patient-level cluster bootstrapping (B=1000). Lower MAE indicates higher accuracy. Source*:* Multimedia Appendix 3 (tab_01_strategy_metrics_ci.csv).

Distributional Fidelity (RQ2)

The ranking changed when the whole distribution was considered (Tables 2 and 3). S4 had the smallest KS distance (0.023), followed by S1 (0.043), S3 (0.051), S2 (0.064), and the Meta Learner (0.067). S4 also placed the most encounters within ±1 point (99.18%) and matched the reference exactly in 87.04% of encounters. All S4 outputs were integers, like the reference.

Exact agreement by reference value showed where strategies lost resolution (Table 5). S4 matched 99.0% of encounters with a reference score of 0% and 71.6% with a score of 1. S3 matched 94.6% and 51.3%, and S2 matched 92.4% and 43.1%. Among encounters with reference scores of 3 to 5, S4 still matched 60% to 64%, compared with 20% to 41% for S2 and S3 (tab_13_exact_agreement_by_value.csv, Multimedia Appendix 5). The continuous strategies spread probability mass between neighboring integers and thus rarely hit a value exactly.

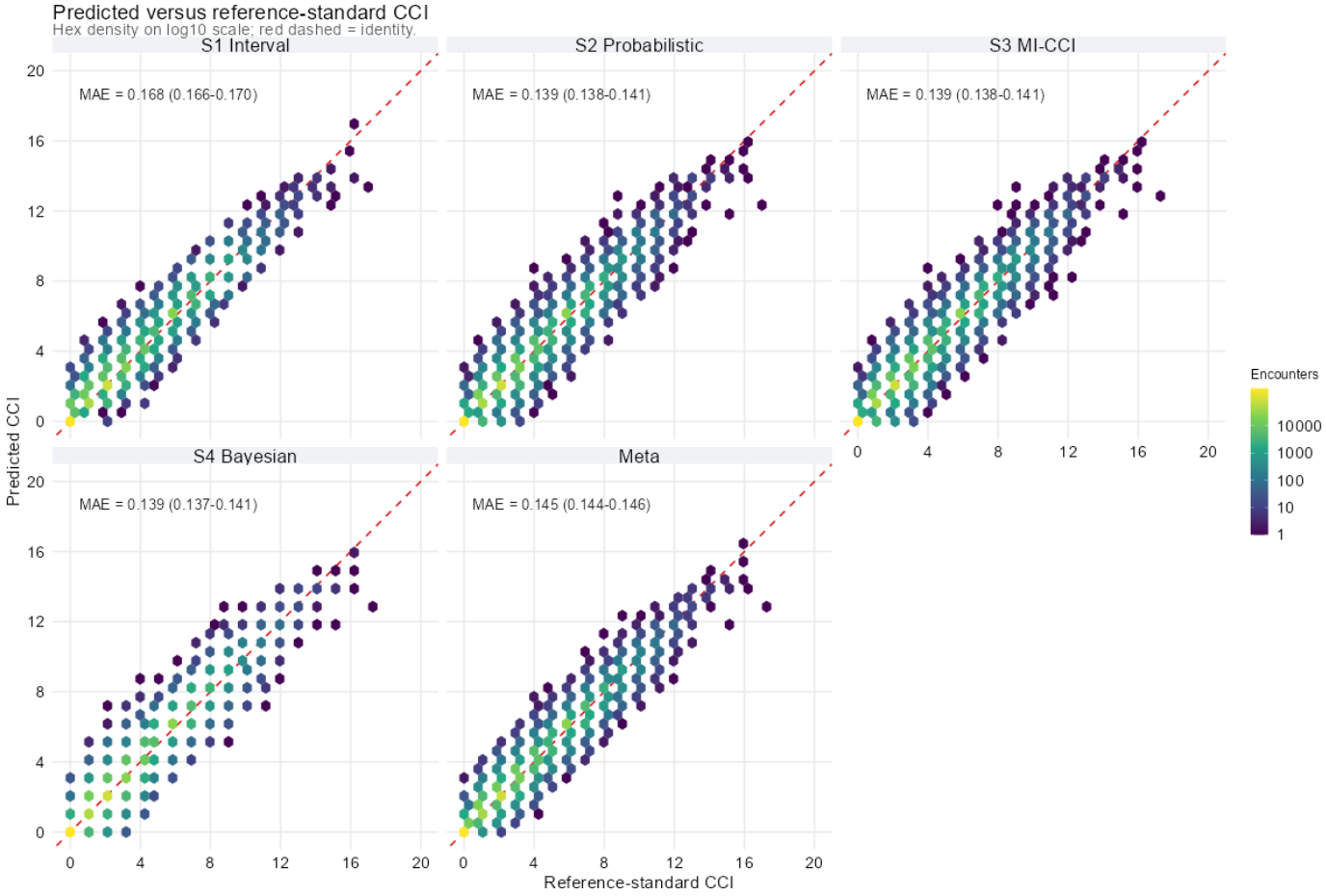
Figure 2 shows the joint distribution of predicted and reference scores. S1 forms a grid of half-integer rows. S2, S3, and the Meta Learner form continuous clouds along the identity line. S4 concentrates on integer rows, which produces its high exact agreement and its higher squared error.

Table 5. Exact agreement with the reference CCI[a] by reference score[b].

Reference CCI	Encounters (n)	S1 (%)	S2 (%)	S3 (%)	S4 (%)	Meta (%)
0	230,152	91.51	92.40	94.57	99.02	91.51
1	66,476	65.03	43.11	51.26	71.63	41.08
2	102,897	81.64	76.78	79.15	88.43	75.31
3	35,954	57.01	35.92	40.71	64.41	33.94
4	17,569	50.36	28.17	33.01	61.61	26.36
5	9159	41.39	19.50	23.53	60.31	18.28

[a]CCI: Charlson Comorbidity Index.
[b]Exact agreement represents the percentage of encounters for which the estimated CCI exactly matches the reference CCI, without rounding or binning. Only reference CCI values from 0 to 5 are presented. S4 achieved the highest exact agreement across all 6 reference scores. Source: Multimedia Appendix 4 (tab_13_exact_agreement_by_value.csv).

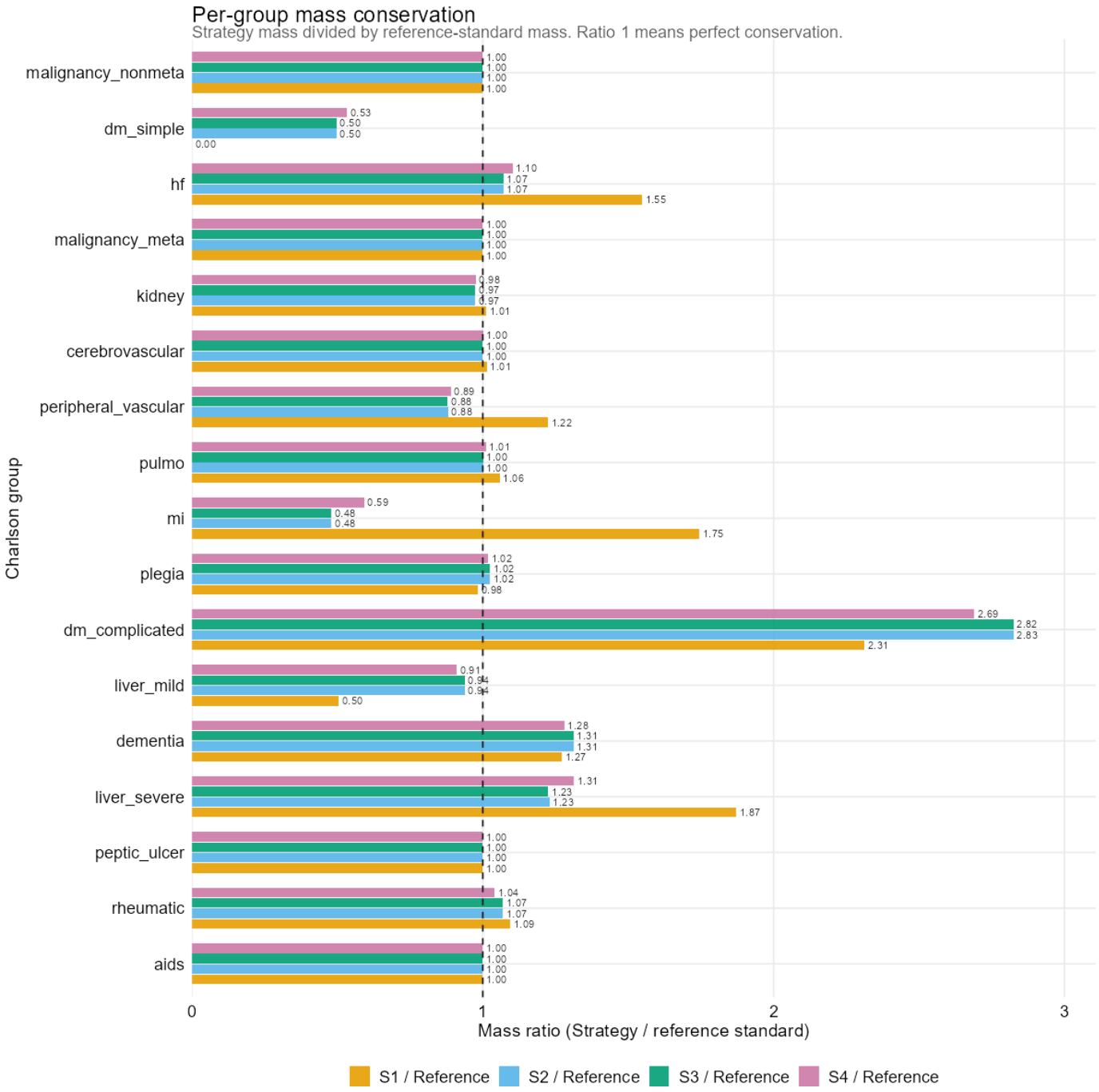
Figure 2. Predicted versus reference standard Charlson Comorbidity Index (CCI) for all 5 strategies (n=507,949). Hexagon color shows encounter density on a log10 scale. The red dashed line marks identity. Strip labels give the mean absolute error (MAE) with its 95% cluster bootstrap CI. S1: strategy 1 (deterministic interval); S2: strategy 2 (exact expected value); S3: strategy 3 (multiple imputation).



Per-Group Mass Conservation

Accurate totals can conceal opposing errors at the group level. The mass conservation ratio locates them. Figure 3 shows this ratio for the 17 Charlson groups (tab_03_qa_group_coverage.csv, Multimedia Appendix 3). Twelve groups were conserved within ±15% under S2, S3, and S4, and 9 under S1. Metastatic and nonmetastatic malignancy, peptic ulcer, and AIDS were conserved exactly by all strategies. For AIDS, this results from the degenerate-prefix rule. Three groups deviated systematically. Complicated diabetes was overattributed (R_g 2.31 to 2.83), and uncomplicated diabetes was underattributed (0.50 to 0.53 under S2 to S4). S1 attributed no mass to uncomplicated diabetes, because every diabetes prefix can also reach the complicated form, which suppresses the uncomplicated form in both bounds. Myocardial infarction was underattributed by S2 to S4 (0.48 to 0.59) and overattributed by S1 (1.75).

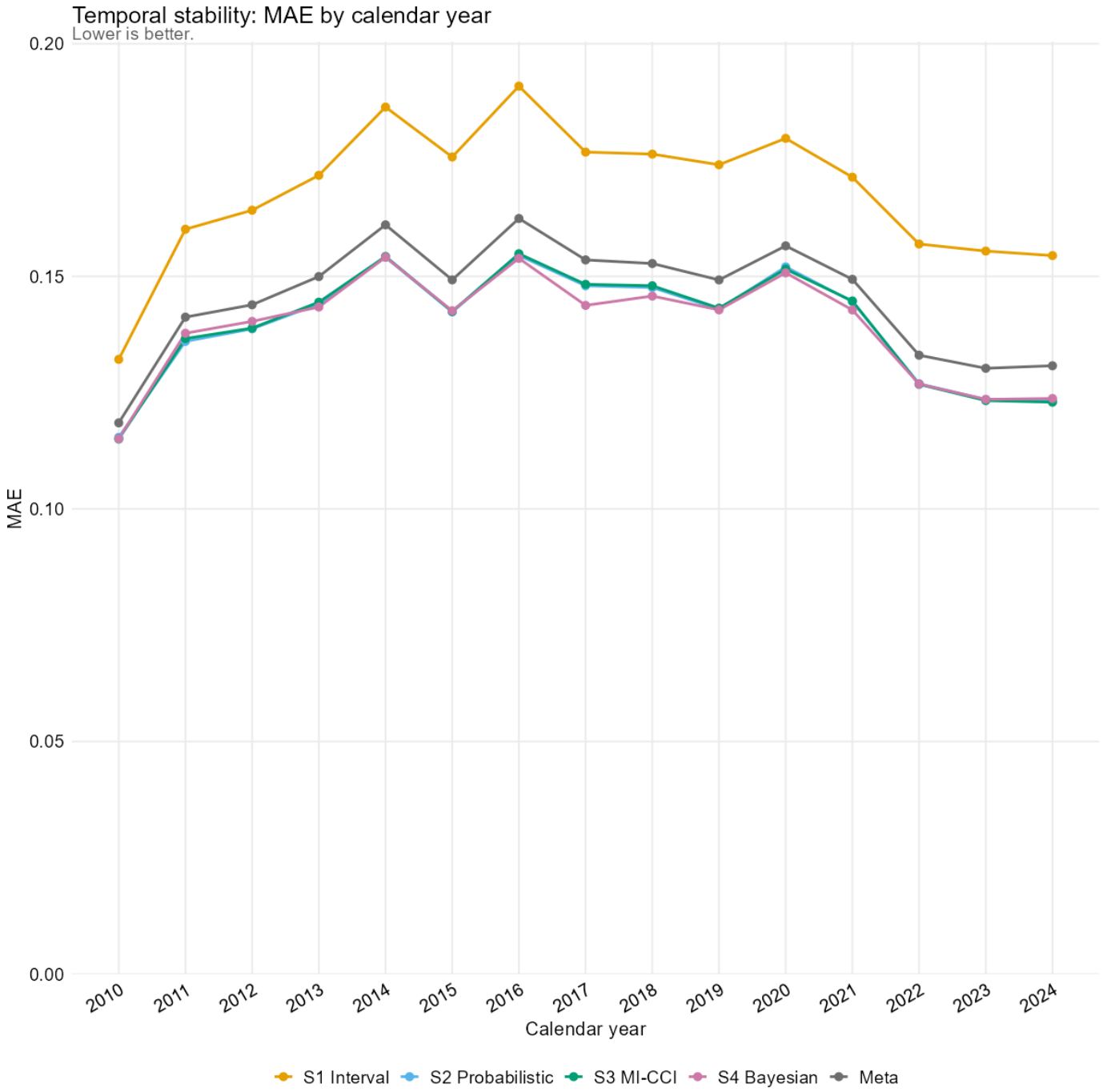
Figure 3. Per-group mass conservation. Bars show the ratio of strategy mass to reference mass for each Charlson group. The dashed line at 1 marks perfect conservation. Values above 1 indicate overattribution, and values below 1 indicate underattribution. Mass is defined in the Methods section (Mass conservation ratio). S1: strategy 1 (deterministic interval); S2: strategy 2 (exact expected value); S3: strategy 3 (multiple imputation); S4: strategy 4 (Dirichlet-smoothed predictive median).



Temporal Stability

Because S1 to S4 fit no parameters on the cohort, error per calendar year probes transportability across 15 years of coding practice. Figure 4 shows the MAE per year. S2 to S4 ranged from about 0.11 to 0.16 and moved in parallel. S1 ranged from about 0.13 to 0.19, and the Meta Learner lay above S2 to S4 in every year. The error rose from 2010 to peak in 2014 and 2016 and fell after 2021. The years 2022 to 2024, closest to the 2023 prior, had among the lowest errors, but 2010 was lower still. Error therefore did not increase steadily with distance from the prior year.

Figure 4. Mean absolute error (MAE) by calendar year for each strategy (2010 to 2024). The 2024 value covers January to September. S1 (strategy 1 [deterministic interval]) to S4 (strategy 4 [Dirichlet-smoothed predictive median]) fit no parameters on the cohort, whereas the Meta Learner is fitted (Methods). CCI: Charlson Comorbidity Index; MAE: mean absolute error; S1: strategy 1 (deterministic interval); S2: strategy 2 (exact expected value); S3: strategy 3 (multiple imputation); S4: strategy 4 (Dirichlet-smoothed predictive median).

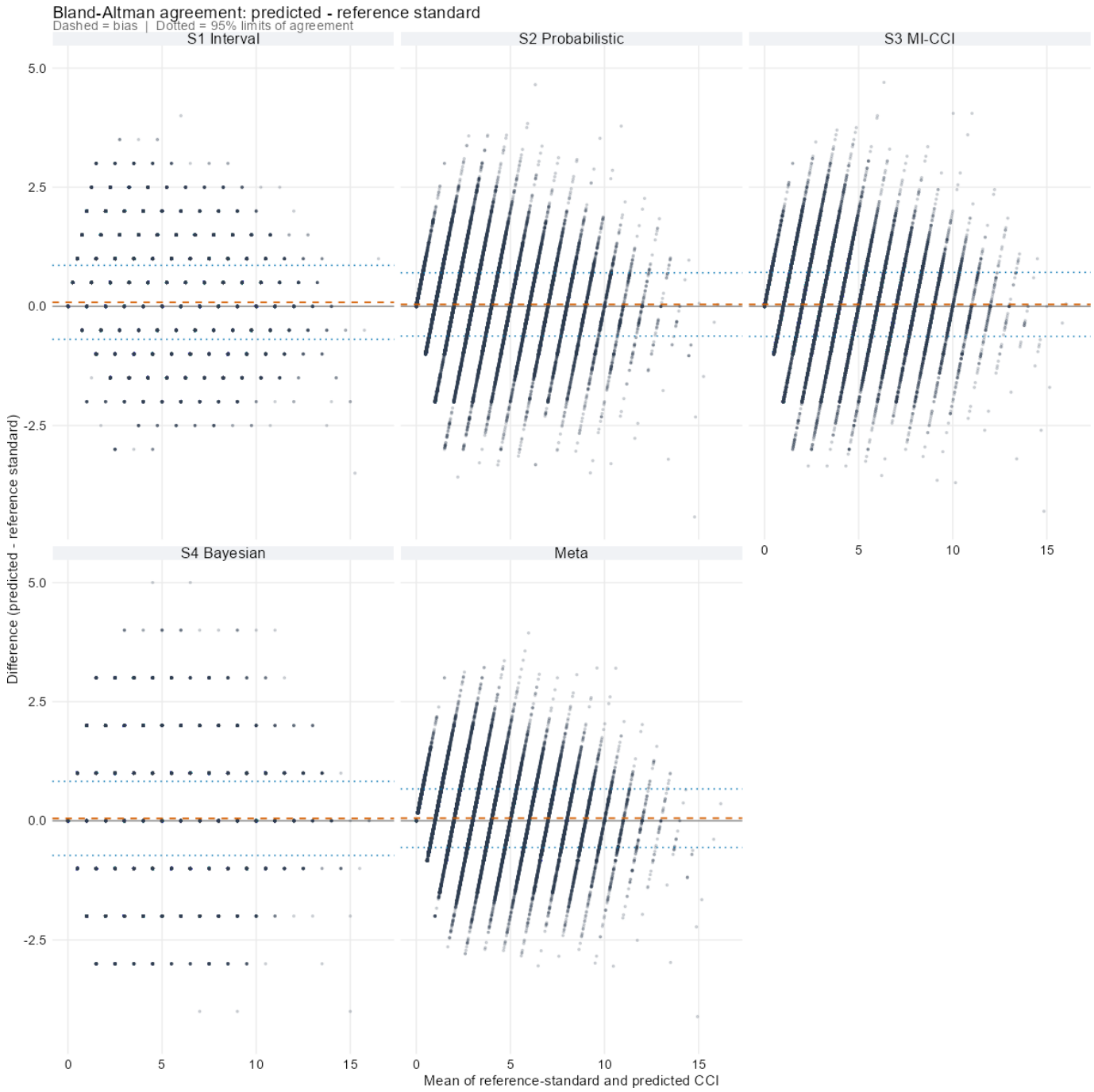


Agreement Across the Score Range

Overview

Bland-Altman plots showed near-zero bias for all strategies (Figure 5, Tables 2 and 3). The narrowest limits of agreement belonged to the Meta Learner (–0.56 to 0.67), followed by S2 and S3. S4 and S1 had wider limits. The plots showed diagonal banding from discrete reference values and heteroscedastic spread at higher scores. The limits therefore summarize agreement only descriptively.

Figure 5. Bland-Altman agreement between predicted and reference standard (Charlson Comorbidity Index) CCI for all 5 strategies (n=507,949). Each point plots the difference (predicted minus reference) against the mean of both. The orange dashed line marks the mean bias. The blue dotted lines mark the 95% limits of agreement (bias ±1.96 SD). Diagonal bands arise from the integer reference. S1: strategy 1 (deterministic interval); S2: strategy 2 (exact expected value); S3: strategy 3 (multiple imputation); S4: strategy 4 (Dirichlet-smoothed predictive median).



Sources of Reconstruction Error (RQ3)

Table 6 summarizes the 3 error source analyses. Under the local oracle prior, about 90% of the MAE of S2 and S3 and 73% of the MAE of S4 remained (tab_pp_01_error_decomposition.csv, Multimedia Appendix 6). S4 therefore gained most from a matched prior. S1 uses no frequencies. Its MAE changed only because the local prior lists a different set of subcodes per prefix. Group-level shifts separated prior mismatch from truncation (tab_pp_02_group_mass_shift.csv). Myocardial infarction, both diabetes groups, and dementia returned within ±15% under the local prior, whereas severe liver disease remained outside. Because the national prior counts principal diagnoses only, this mismatch reflects diagnosis role as well as case mix. It may contribute to the underattribution of myocardial infarction and the overattribution of complicated diabetes. Among Charlson-relevant prefixes, E11 and I25 showed the largest exposure-weighted divergence between the two priors (tab_pp_03_prior_divergence.csv).

Thirty-nine of the 347 degenerate prefixes in the cohort were Charlson-scoring, for example C61, F03, and C20 (tab_sg_02_degenerate_prefixes.csv and

tab_sg_03_cohort_exposure.csv, Multimedia Appendix 7). Silent dropping would have discarded these diagnoses and changed the scores of 35,156 encounters (tab_sg_04_score_impact.csv). Four of the 10 degenerate prefixes with the most encounters were secondary-only codes (U99, B96, B95, and U69), and asterisk codes such as D63, N08, and G63 also lacked Destatis children (tab_sg_02_degenerate_prefixes.csv). This pattern is consistent with a prior restricted to principal diagnoses. Among Charlson groups, dementia had the largest silent gap, with 10 of 22 four-character codes absent from Destatis (tab_sg_01_catalogue_coverage.csv).

Monte Carlo settings mattered for S4 but hardly for S3 (Multimedia Appendix 4). The S4 MAE depended mainly on the aggregator and on the prior strength (tab_hp_02_s4_sweep.csv and fig_hp_01_sensitivity.png). At default settings, S4 and the best S3 setting were tied (tab_hp_03_paired_vs_best_s3.csv). S4 outperformed S3 after Bonferroni correction only with the median aggregator and $\alpha_0=1$. The ranking of the 2 strategies therefore depends on the S4 setting (tab_hp_04_verdict.csv). The median aggregator kept the KS distance of S4 low in every cell.

The local oracle prior is estimated in sample from the cohort's own full-length codes and pooled over 2010-2024. Its error is an optimistic benchmark, not a deployable alternative. Residual and mismatch shares are percentages of the national-prior MAE. For S1, which uses no frequencies, the decomposition is not meaningful (mismatch –2.5%). Silent-drop comparisons used the marginal prior. The sensitivity subsample comprised 10,000 patients with 20,973 encounters. The best S3 setting was $M=100$. Source files: tab_pp_01_error_decomposition.csv and tab_pp_02_group_mass_shift.csv (Multimedia Appendix 6), tab_05_destatis_vs_catalogue_global.csv and tab_sg_01_catalogue_coverage.csv to tab_sg_04_score_impact.csv (Multimedia Appendix 7), and tab_hp_01_s3_sweep.csv to tab_hp_04_verdict.csv (Multimedia Appendix 4). Produced by analysis scripts 05 to 08 (Multimedia Appendix 2).

Table 6. Sources of reconstruction error: prior provenance, prefixes without Destatis subcodes, and Monte Carlo settings (RQ3)[a].

Quantity	Result
Prior provenance (full cohort)	
MAE[b] with national prior versus local oracle prior	S1[c] 0.168 versus 0.172, S2[d] 0.139 versus 0.125, S3[e] 0.139 versus 0.126, S4[f] 0.139 versus 0.102
Share of MAE remaining under the oracle prior	S2 89.6%, S3 90.0%, S4 73.1%
Share of MAE caused by prior mismatch	S2 10.4%, S3 10.0%, S4 26.9%
Mass ratio of S3, national versus local prior	Myocardial infarction 0.48 versus 0.97, uncomplicated diabetes 0.50 versus 0.96, complicated diabetes 2.82 versus 1.14, dementia 1.31 versus 1.08, severe liver disease 1.23 versus 1.19
Prefixes without Destatis subcodes	
4-character catalog codes absent from Destatis	1622 of 9306 (17.4%)
Degenerate prefixes in the cohort	347 of 1601 (39 Charlson-scoring)
Encounters with a scoring degenerate prefix	55,248 (10.9%)
MAE with silent dropping versus self-fallback	S1 0.277 versus 0.168, S2 0.252 versus 0.141
Monte Carlo settings (subsample)	
S3 MAE across $M=5$ to 100	0.1406 to 0.1396 (range 0.0010)
S4 MAE across 30 cells	0.1328 to 0.1472 (range 0.0144)
Default S4 minus best S3 (95% CI)	–0.0018 (–0.0044 to 0.0008), tie
Bonferroni verdicts over 30 cells	S4 better in 4, S3 better in 11, tie in 15
S4 KS[g] distance, median versus mean aggregator	0.019 to 0.026 versus 0.044 to 0.064

[a]RQ: research question.
[b]MAE: mean absolute error.
[c]S1: strategy 1 (deterministic interval).
[d]S2: strategy 2 (exact expected value).
[e]S3: strategy 3 (multiple imputation).
[f]S4: strategy 4 (Dirichlet-smoothed predictive median).
[g]KS: Kolmogorov-Smirnov.

Meta Learner Composition

The fitted weights explain why the ensemble lowers squared error but not MAE. The NNLS weights were 0.630 for S2, 0.331 for S1, 0.039 for S3, and 0 for S4 (Tables 2 and 3, meta_weights.csv, Multimedia Appendix 3). S2 had the lowest cross-validated risk (0.116). S3 carries the signal of S2 plus Monte Carlo noise and received almost no weight. S4 draws from the same prior, and its integer output adds squared error (risk 0.160) without complementary signal. S1 received substantial weight despite its high risk (0.164), because its frequency-free midpoint errs through a different mechanism. The composition was stable. Across the 10 outer folds of the nested cross-validation, the largest weight range was 0.015 (S2), and S4 received 0 in every fold (tab_mo_01_fold_weights.csv, Multimedia Appendix 4). Overall, the ensemble reduced the MSE to 0.101 (95% CI 0.0996-0.1025), but not the MAE.

Subgroup Behavior (RQ4, Exploratory)

Table 7 summarizes the subgroup comparisons. Bonferroni correction removed 1 of 11 decisive chapters, 3 of 12 decisive Charlson groups, and 3 of 26 decisive patterns (tab_mc_04_summary.csv, Multimedia Appendix 8). By encounter count, S4 led most Bonferroni-decisive chapters and patterns, including circulatory and digestive diseases. S1 led neoplasms, endocrine diseases, and 8 patterns containing uncomplicated diabetes. Per-stratum metrics and winners are provided in tab_07_stratified_chapter.csv to tab_12_winners_pattern.csv (Multimedia Appendix 5), and stratified figures in Multimedia Appendix 9.

A stratum is decisive when the paired cluster bootstrap CI of the MAE difference between the best and the second-best strategy excludes zero. Bonferroni levels were 0.9975 (chapters), 0.99706 (groups), and 0.99833 (patterns). Encounters can belong to several Charlson groups. Source files: tab_mc_01_chapter_bonferroni.csv, tab_mc_02_group_bonferroni.csv, tab_mc_03_pattern_bonferroni.csv, and tab_mc_04_summary.csv (Multimedia Appendix 8), and tab_10_winners_chapter.csv to tab_12_winners_pattern.csv (Multimedia Appendix 5). Produced by analysis/09_multiplicity_correction.R (Multimedia Appendix 2).

Table 7. Exploratory subgroup comparisons with Bonferroni correction (RQ4)[a].

Family	Strata (K)	Decisive at 95%	Decisive after Bonferroni	Bonferroni-decisive leaders, strata (encounters)	Lost after correction
ICD-10[b] chapter	20	11	10	• S4[c]: 5 (182,717) • S1[d]: 4 (158,978) • Meta: 1 (45,328)	I Infectious
Active Charlson group	17	12	9	• S4: 5 (88,945) • S1: 3 (112,789) • Meta: 1 (97,265)	Cerebrovascular, heart failure, peptic ulcer
Multimorbidity pattern	30	26	23	• S4: 11 (140,028) • S1: 12 (48,408)	Kidney, kidney with nonmetastatic malignancy, and heart failure with pulmonary disease

[a]RQ: research question.
[b]*ICD-10*: International Statistical Classification of Diseases.
[c]S4: strategy 4 (Dirichlet-smoothed predictive median).
[d]S1: strategy 1 (deterministic interval).

Discussion

Principal Findings

In this single-center cohort, frequency-informed estimation reconstructed the CCI from three-character *ICD-10-GM* codes with an MAE of about 0.14 points (S2: 0.1394, 95% CI 0.1379-0.1408; S3: 0.1395, 95% CI 0.1380-0.1408; S4: 0.1390, 95% CI 0.1374-0.1405). Three findings stand out. First, S2, S3, and S4 were nearly identical on MAE. S2 is the exact expectation that S3 approximates, and in paired testing on a subsample of 10,000 patients, S4 at its default setting did not differ from the best S3 setting. Thus, no MAE advantage was demonstrated for default S4 over the best S3 setting. Second, the strategies differed in which property they preserved. S2 and S3 minimized squared error among single strategies, whereas S4 preserved the integer distribution of the reference. The Meta Learner achieved the lowest squared error but not the lowest MAE, also under nested cross-validation. Third, most of the error persisted even under a cohort-matched prior. Prior mismatch added 10% to 27%, with the largest group-level shifts in diabetes, myocardial infarction, and dementia. This mismatch reflects case mix and the restriction of the national before principal diagnoses.

Comparison to Prior Work

CCI mappings have been validated across countries [48], and a German-specific mapping is now available [14]. All of them assume native subcode specificity. The Stausberg index [20] handles 3-digit codes by defining a new instrument, which limits comparability with the international CCI literature. *miCCI* keeps the Quan definition and the CCI scale. Reconstructed scores therefore remain comparable with published CCI norms [6]. Simard et al [3] found marked heterogeneity among *ICD*-based multimorbidity measures. Keeping an established instrument computable avoids adding a further variant.

Multiple imputation is standard for missing values in health records [31,49]. S3 applies it to a different kind of missingness, in which the mechanism is deterministic and known, and the hidden suffix follows an external distribution. Our results show that, for an additive score with known suppression rules, this imputation has a closed form. S2 computes the same expectation without sampling. This observation matters beyond the CCI. Additive indices with hierarchical rules of this kind can be completed analytically under the same independence assumption. Simulation is needed only when a distributional summary such as the median is required.

Privacy research treats diagnosis codes as quasi-identifiers [17,50]. Our prior work showed that k-anonymization can alter distributional shape and rare-diagnosis representation without being flagged by internal loss metrics [19]. The current study adds a complementary perspective. It quantifies the residual analytical cost of one generalization step after reconstruction with a public prior.

The *comorbidity* [21] and *icd* [22] R packages require full-length codes. *miCCI* acts upstream of them and returns a reconstructed score for standard analyses. The per-group mass conservation analysis adds a further element. Conventional validations report scalar agreement [5,48]. Mass conservation showed that accurate aggregate scores can hide group-level shifts, such as the diabetes redistribution. The oracle analysis attributed these shifts largely to prior mismatch in case mix and diagnosis role.

Recommendations and Practice Insights

Strategy choice should follow the role of the CCI in the downstream analysis. S2 is recommended when encounter-level accuracy matters most, for example when the CCI enters a regression model as a covariate. It is deterministic, fast, and free of Monte Carlo noise. S3 gives nearly identical results but is slower. S4 is recommended when the score distribution matters, for example for cohort description or for pooling score distributions across sites. Its behavior at fixed thresholds such as CCI ≥2 has not yet been evaluated. S1 remains useful when frequencies are unavailable or explicit bounds are required. The Meta Learner requires a local calibration set with full-length codes and is preferable only when squared error is the target.

For data integration centers, 3 practical points follow from the findings. Reconstruction can be performed locally before release, with the public Destatis table and the open-source package as the only inputs. In this cohort, all 5 strategies were completed for 507,949 encounters in 8.2 hours on one workstation. Prefixes without 4-character children must be retained rather than discarded, because they occurred in 10.9% (55,248/507,949) of encounters and their omission raised the MAE by about 0.11 points (S1: 0.168 to 0.277; S2: 0.141 to 0.252). A locally derived prior, where available, is expected to remove up to one quarter of the remaining error, since prior mismatch accounted for 10% to 27% of the MAE (Table 6). Finally, the S1 interval width bounds, for each encounter, the score range compatible with its truncated codes, and it can be released alongside the score as a per-encounter quality indicator. What these steps do not yet establish is whether reconstructed scores leave risk-adjusted outcome models unchanged. That question defines the next study. Textbox 1 shows a minimal example.

Textbox 1. Minimal example for one truncated encounter (type 2 diabetes mellitus, chronic kidney disease, and essential hypertension) using the 4 base strategies of *miCCI*. Outputs are those documented in README.md (Multimedia Appendix 2) for the national marginal prior and the Destatis snapshot bundled with version 1.0.0. They may shift slightly if a newer Destatis release is loaded.

```
remotes::install_github("gaetankamdje-wabo/miCCI@v1.0.0")
library(miCCI)
quan_map <- load_quan_map()
cache <- precompute_lookups(load_destatis(), quan_map)
codes <- c("E11," "N18," "I10")
cci_interval(codes, quan_map, cache)$cci_mid # S1: 3
cci_probabilistic(codes, quan_map, cache)$e_cci # S2: 3.586574
cci_mi(codes, quan_map, cache, m=20L, seed =42L)$mi_cci # S3: 3.6
cci_bayesian(codes, quan_map, cache, n_draws =25L,
   alpha_0=10, seed =42L)$posterior_median # S4: 4
# I10 lies outside the Quan mapping and is ignored by every strategy.
```

Limitations

First, the evaluation used one center, and the reference derives from the same center's codes. The temporal analysis and the oracle decomposition partly address transportability. Error profiles may still differ at other sites, and multisite validation is required.

Second, a static 2023 prior was applied to data from 2010 to 2024. Coding practice changes over time [45], so older years may carry a larger prior mismatch. The yearly errors did not increase steadily with distance from 2023, which suggests that chronology is not the dominant driver. Year-matched Destatis releases would evaluate this directly.

Third, the national prior does not match the case mix of a tertiary center. It also counts principal diagnoses only, whereas Charlson conditions are mostly coded as secondary diagnoses. The oracle analysis estimated both effects together at 10% to 27% of the MAE. Because the oracle is estimated in sample and pooled over years, this share is approximate. Per-group inference remains biased for diabetes, myocardial infarction, and dementia under the national prior. Regional or institutional priors built from all diagnosis positions could reduce this bias.

Fourth, degenerate prefixes are matched in both directions. This choice avoids silent loss of valid codes but may overactivate groups whose patterns cover only part of a degenerate prefix. It affects at most 10.9% (55,248/507,949) of encounters with a scoring degenerate prefix.

Fifth, S2 to S4 assume that coded positions are completed independently. Within-encounter dependence between subcodes, for example between diabetes and kidney complications, is not modeled. A conditional prior could capture it.

Sixth, the Meta Learner was validated by nested cross-validation within one center only. Its weights are site-specific, so transfer to another site requires refitting on a local calibration set with full-length codes.

Seventh, hyperparameter sensitivity was run on a subsample, and the main analysis used one random seed. The subsample reproduced full-cohort accuracy, and the paired design controls noise between settings. Multiseed runs would quantify Monte Carlo variance directly. S4 also applied one fixed prior strength ($\alpha_0=10$) to every prefix, although $\alpha_0=1$ gave the lowest MAE in the sensitivity grid (Multimedia Appendix 4). Moreover, the package does not enforce an odd n_draws. With an even value, the median of the 2 middle scores can be a half-integer, so S4 may lose its integer support when users override the default of 25 draws.

Eighth, the headline MAE includes the 45.31% (230,152/507,949) of encounters with a reference score of zero, many with unambiguous prefixes. It therefore understates the error among ambiguous, higher-score encounters. The value-specific exact agreement in Table 5 partly addresses this. In addition, sex was not available, and the prior pools the sex-specific rows of Table 23131-01. German-specific U-codes are outside the Quan set, and the CCI is not validated for children. Finally, the Bland-Altman limits in Tables 2 and 3 were derived from summary statistics under a normal approximation, which describes agreement only coarsely for discrete, heteroscedastic scores.

Future Directions

The next step is a downstream utility study. It will measure agreement at the thresholds CCI ≥2 and CCI ≥3, and the change in adjusted length-of-stay and mortality models when reconstructed scores replace reference scores. Further work includes ensembles trained on absolute error, year-matched and regional priors, and federated multisite validation. Planned package refinements should address the technical limitations highlighted above. They will include a parity check or warning for n_draws in S4, an option in load_destatis() to retain sex stratification, an evaluation of nonparametric limits of agreement from the 2.5th and 97.5th percentiles of encounter-level differences, and investigation of a prior strength scaled to the national frequency of each prefix (an empirical Bayes approach). Such validation needs only frequency tables and aggregate outputs. However, the reconstruction engine is instrument-agnostic, so the Elixhauser measures [51] are a natural extension.

Pediatric comorbidity offers a promising niche. Dedicated pediatric indices outperform adult scores in children [25,26], and childhood multimorbidity lacks standardized measurement [27]. Because *miCCI* needs only a code mapping and a frequency table, the same engine could reconstruct pediatric indices from truncated codes. The 5 pediatric Destatis age bands would supply age-specific priors where adult frequencies fit least. Pseudonymized patient identifiers would allow tracking of reconstructed comorbidity across the transition from pediatric to adult care. Pediatric cohorts are small at single sites, so federated pooling of frequency tables would be particularly valuable. Finally, the S1 interval width could serve as a data quality flag that tells users of released pediatric data how much comorbidity information truncation has removed.

Conclusions

In this single-center cohort of 507,949 inpatient encounters, frequency-informed strategies reconstructed the CCI from 3-character *ICD-10-GM* codes with an MAE of about 0.14 points (S2: 0.1394, 95% CI 0.1379-0.1408; S3: 0.1395, 95% CI 0.1380-0.1408; S4: 0.1390, 95% CI 0.1374-0.1405). The expected-value estimator (S2) and multiple imputation (S3) were nearly identical, because S3 approximates the expectation that S2 computes exactly. The Dirichlet-smoothed median (S4) matched their MAE and best preserved the integer distribution of the reference. Most of the remaining error persisted under a cohort-matched prior. The open-source *miCCI* package makes these strategies available for 3-character truncated *ICD-10-GM* data. Their value for clinical research is a potential application that requires multisite confirmation.

Acknowledgments

The authors thank the Data Integration Center of University Hospital Mannheim for the pseudonymized data extraction. We acknowledge financial support for the publication fee from Heidelberg University.

Generative AI Use

The authors disclose the use of generative AI tools in the analytical workflow and manuscript preparation. Following the GAIDeT taxonomy [52], Manus 1.6 was used to provide comments on, and assist with the organization and optimization of, manually generated R analysis scripts during revision of the manuscript. Grammarly was used to support grammatical and linguistic editing. All uses of these tools occurred under direct human supervision. Generative AI was not used to conceptualize the study, select or interpret the results, or formulate scientific claims and conclusions. The authors reviewed and verified all AI-assisted outputs and take full responsibility for the accuracy, integrity, and final content of the manuscript.

Funding

This publication was partially supported by the German Federal Ministry of Research, Technology and Space (Bundesministerium für Forschung, Technologie und Raumfahrt, BMFTR) through the Network of University Medicine 3.0 (NUM 3.0), grant number 01KX2524.

Data Availability

Patient-level data cannot be shared publicly because of data protection requirements. Access may be considered individually, subject to approval by the ethics committee and the use and access committees. The *miCCI* source code is available under the MIT license as the tagged release v1.0.0 [24]. The development build labeled 1.1.0 that produced the results

(run_manifest.json, Multimedia Appendix 3) is computationally equivalent to this release. Users can replace the Destatis local file path with the official Destatis URL [23] to load an updated release (miCCI/R/02_destatis.R, Multimedia Appendix 1). All aggregate outputs underlying the tables and figures are provided in Multimedia Appendices 1 to 9. In these outputs, variable names containing gold (for example, cci_gold and gold_mass) and the row label Gold Standard denote the reference standard.

Authors' Contributions

Conceptualization: GKW, FS
Data curation: GKW, MJL
Formal analysis: GKW
Funding acquisition: TG, FS
Investigation: GKW
Methodology: GKW (study methodology and CCI mapping), PPS (CCI mapping), FS (CCI mapping)
Project administration: FS
Software: GKW (lead; software design, implementation, and technical leadership)
Supervision: FS
Validation: GKW, FS, MH
Writing – original draft: GKW
Writing – review & editing: GKW, NS, MJL, PPS, MH, TG, FS

Conflicts of Interest

None declared.

Multimedia Appendix 1

Cohort selection and descriptive characteristics, including stepwise exclusions, age distribution, diagnosis counts, length of stay, and annual encounter volumes, presented in tables and figures.
[ZIP File (ZIP archive File), 24235 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Source code and implementation materials for the *miCCI* R package, including computational strategies for Charlson Comorbidity Index (CCI) reconstruction, analysis and orchestration scripts, unit tests, *International Statistical Classification of Diseases, 10th Revision, German Modification* (*ICD-10-GM*)-to-Charlson mappings, and the *ICD-10-GM* 2024 catalog.
[ZIP File (ZIP archive File), 411 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Performance comparisons, statistical significance assessments, and nested cross-validation of the Meta Learner with fold-specific weights, performance metrics, and 95% CIs.
[ZIP File (ZIP archive File), 7 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Hyperparameter sensitivity analyses of the S3 and S4 computational strategies and nested cross-validation of the Meta Learner, including performance comparisons, paired bootstrap analyses, fold-specific ensemble weights, and 95% CIs.
[ZIP File (ZIP archive File), 41 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Stratified performance evaluation of the five computational strategies across (*International Statistical Classification of Diseases, 10th Revision, German Modification) ICD-10-GM* chapters, Charlson comorbidity groups, and multimorbidity patterns, including subgroup-specific strategy comparisons, best-performing strategies, and exact agreement with the reference Charlson Comorbidity Index (CCI).
[ZIP File (ZIP archive File), 41 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Prior provenance and sensitivity analyses comparing national and local diagnosis-code frequency distributions, including reconstruction error decomposition, Charlson comorbidity group-level contribution shifts, and divergence between population-based and local priors.
[ZIP File (ZIP archive File), 15 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Comparison of Destatis hospital diagnosis frequencies (table 23131-01) with the *International Statistical Classification of Diseases, 10th Revision, German Modification* (*ICD-10-GM)* 2024 catalog, including diagnosis-code coverage, discrepancies in code availability, degenerate prefixes, cohort exposure to missing subcodes, and the impact of these coding gaps on Charlson Comorbidity Index (CCI) reconstruction accuracy.
[ZIP File (ZIP archive File), 192 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Bonferroni-adjusted paired bootstrap analyses comparing the 5 computational strategies across *International Statistical Classification of Diseases, 10th Revision, German Modification* (*ICD-10-GM)* chapters, Charlson comorbidity groups, and multimorbidity patterns, including multiplicity-adjusted CIs and summaries of statistically significant subgroup differences.
[ZIP File (ZIP archive File), 4 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Stratified mean absolute error (MAE) comparisons of the 5 computational strategies across *International Statistical Classification of Diseases, 10th Revision, German Modification* (*ICD-10-GM)* chapters, Charlson comorbidity groups, and multimorbidity patterns, presented as supplementary figures.
[ZIP File (ZIP archive File), 587 KB-Multimedia Appendix 9]

References

1. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
2. Sarfati D, Koczwara B, Jackson C. The impact of comorbidity on cancer and its treatment. CA Cancer J Clin. Jul 2016;66(4):337-350. [doi: 10.3322/caac.21342] [Medline: 26891458]
3. Simard M, Rahme E, Calfat AC, Sirois C. Multimorbidity measures from health administrative data using ICD system codes: a systematic review. Pharmacoepidemiol Drug Saf. Jan 2022;31(1):1-12. [doi: 10.1002/pds.5368] [Medline: 34623723]
4. Sharabiani MTA, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. Med Care. Dec 2012;50(12):1109-1118. [doi: 10.1097/MLR.0b013e31825f64d0] [Medline: 22929993]
5. Yurkovich M, Avina-Zubieta JA, Thomas J, Gorenchtein M, Lacaille D. A systematic review identifies valid comorbidity indices derived from administrative health data. J Clin Epidemiol. Jan 2015;68(1):3-14. [doi: 10.1016/j.jclinepi.2014.09.010] [Medline: 25441702]
6. Boogaerts G, Loyd C, Zhang Y, Kennedy RE, Brown CJ. National norms for the Elixhauser and Charlson comorbidity indexes among hospitalized adults. J Gerontol A Biol Sci Med Sci. Feb 24, 2023;78(2):365-372. [doi: 10.1093/gerona/glac087] [Medline: 35426436]
7. Arrigo RT, Kalanithi P, Cheng I, et al. Charlson score is a robust predictor of 30-day complications following spinal metastasis surgery. Spine. Sep 1, 2011;36(19):E1274-E1280. [doi: 10.1097/BRS.0b013e318206cda3] [Medline: 21358481]
8. Lakomkin N, Zuckerman SL, Stannard B, et al. Preoperative risk stratification in spine tumor surgery: a comparison of the modified Charlson index, frailty index, and ASA score. Spine (Phila Pa 1976). Jul 1, 2019;44(13):E782-E787. [doi: 10.1097/BRS.0000000000002970] [Medline: 31205174]
9. Toson B, Harvey LA, Close JCT. The ICD-10 Charlson comorbidity index predicted mortality but not resource utilization following hip fracture. J Clin Epidemiol. Jan 2015;68(1):44-51. [doi: 10.1016/j.jclinepi.2014.09.017] [Medline: 25447352]
10. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol. Jun 1992;45(6):613-619. [doi: 10.1016/0895-4356(92)90133-8] [Medline: 1607900]
11. Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J Clin Epidemiol. Oct 1993;46(10):1075-1079. [doi: 10.1016/0895-4356(93)90103-8] [Medline: 8410092]
12. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. Nov 2005;43(11):1130-1139. [doi: 10.1097/01.mlr.0000182534.19832.83] [Medline: 16224307]
13. Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. J Clin Epidemiol. Dec 2004;57(12):1288-1294. [doi: 10.1016/j.jclinepi.2004.03.012] [Medline: 15617955]

14. Sokołowski PP, Hagmann M, Maros ME, Kamdje Wabo G, Meerjanssen JM, Siegel F. Developing country-specific Charlson comorbidity index mappings for use with German administrative data: methodological comparative study. JMIR Med Inform. Sep 3, 2026;14:e93923. [doi: 10.2196/93923] [Medline: 42593349]
15. Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. ARX--a comprehensive tool for anonymizing biomedical data. AMIA Annu Symp Proc. 2014;2014:984-993. [Medline: 25954407]
16. Ganslandt T, Mate S, Helbing K, Sax U, Prokosch HU. Unlocking data for clinical research - the German i2b2 experience. Appl Clin Inform. 2011;2(1):116-127. [doi: 10.4338/ACI-2010-09-CR-0051] [Medline: 23616864]
17. Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc. 2010;17(3):322-327. [doi: 10.1136/jamia.2009.002725] [Medline: 20442151]
18. Kamdje Wabo G, Prasser F, Gierend K, Siegel F, Ganslandt T. Data quality- and utility-compliant anonymization of common data model-harmonized electronic health record data: protocol for a scoping review. JMIR Res Protoc. Aug 11, 2023;12:e46471. [doi: 10.2196/46471] [Medline: 37566443]
19. Kamdje Wabo G, Sokolowski PP, Jannesari Ladani M, Hagmann M, Ganslandt T, Siegel F. Quantifying the impact of anonymization-induced clinical data quality loss: methodological quantitative case study using primary diagnosis codes and hospital length of stay. JMIR Med Inform. Sep 11, 2026;14:e97661. [doi: 10.2196/97661] [Medline: 42726783]
20. Stausberg J, Hagn S. New morbidity and comorbidity scores based on the structure of the ICD-10. PLoS ONE. 2015;10(12):e0143365. [doi: 10.1371/journal.pone.0143365] [Medline: 26656501]
21. Gasparini A. Comorbidity: an R package for computing comorbidity scores. JOSS. 2018;3(23):648. [doi: 10.21105/joss.00648]
22. Wasey JO. ICD: comorbidity calculations and tools for ICD-9 and ICD-10 codes. R package. 2019. URL: https://CRAN.R-project.org/package=icd [Accessed 2026-09-25]
23. Diagnosedaten der krankenhäuser (krankenhausdiagnosestatistik) 23131-01. Statistisches Bundesamt (Destatis); 2024. URL: https://www.destatis.de/static/DE/dokumente/5231301237015_SB.xlsx [Accessed 2026-10-27]
24. Kamdje Wabo G, Santhanam N, Jannesari Ladani M, Hagmann M, Sokolowski PP, Ganslandt T, et al. MiCCI: Charlson comorbidity index estimation from truncated ICD-10-GM codes. GitHub. 2026. URL: https://github.com/gaetankamdje-wabo/miCCI/tree/v1.0.0 [Accessed 2026-11-14]
25. Sun JW, Bourgeois FT, Haneuse S, et al. Development and validation of a pediatric comorbidity index. Am J Epidemiol. May 4, 2021;190(5):918-927. [doi: 10.1093/aje/kwaa244] [Medline: 33124649]
26. Ikeda Kurakawa K, Okada A, Konishi T, et al. Children comorbidity score, a simple predictor for in-hospital mortality: a nationwide inpatient database study in Japan. JMA J. Apr 28, 2025;8(2):568-579. [doi: 10.31662/jmaj.2024-0333] [Medline: 40415997]
27. Pradhan SK, Murmu J, Nayak U, et al. The magnitude of multimorbidity in childhood: a global systematic review. Syst Rev. Feb 17, 2026;15(1):67. [doi: 10.1186/s13643-026-03087-z] [Medline: 41703588]
28. ICD-10-GM: internationale statistische klassifikation der krankheiten und verwandter gesundheitsprobleme, 10. Bundesinstitut für Arzneimittel und Medizinprodukte. URL: https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-10-GM/_node.html [Accessed 2026-09-09]
29. von Hippel PT. How many imputations do you need? A two-stage calculation using a quadratic rule. Sociol Methods Res. Aug 2020;49(3):699-718. [doi: 10.1177/0049124117747303] [Medline: 39211325]
30. Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons; 1987. [doi: 10.1002/9780470316696]
31. van Buuren S. Flexible Imputation of Missing Data. 2nd ed. CRC Press; 2018. [doi: 10.1201/9780429492259]
32. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. 3rd ed. CRC Press; 2013. [doi: 10.1201/b16018]
33. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007;6:Article25. [doi: 10.2202/1544-6115.1309] [Medline: 17910531]
34. Polley EC, LeDell E, Kennedy C, Laan MJ. SuperLearner: super learner prediction. R package version 2. 2026:42. [doi: 10.32614/CRAN.package.SuperLearner]
35. Phillips RV, van der Laan MJ, Lee H, Gruber S. Practical considerations for specifying a super learner. Int J Epidemiol. Aug 2, 2023;52(4):1276-1285. [doi: 10.1093/ije/dyad023] [Medline: 36905602]
36. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Nature; 2009. [doi: 10.1007/978-0-387-84858-7]
37. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. Feb 23, 2006;7:91. [doi: 10.1186/1471-2105-7-91] [Medline: 16504092]
38. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. Feb 8, 1986;1(8476):307-310. [doi: 10.1016/S0140-6736(86)90837-8] [Medline: 2868172]

39. Willmott CJ, Matsuura K. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in assessing average model performance. Clim Res. 2005;30(1):79-82. [doi: 10.3354/cr030079]
40. Bioanalytical method validation and study sample analysis. ICH harmonisation for better health; 2022. URL: https://database.ich.org/sites/default/files/M10_Guideline_Step4_2022_0524.pdf [Accessed 2026-10-27]
41. Field CA, Welsh AH. Bootstrapping clustered data. J R Stat Soc Series B Stat Methodol. Jun 1, 2007;69(3):369-390. [doi: 10.1111/j.1467-9868.2007.00593.x]
42. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall/CRC; 1994. [doi: 10.1201/9780429246593]
43. Davison AC, Hinkley DV. Bootstrap Methods and Their Application. Cambridge University Press; 1997. [doi: 10.1017/CBO9780511802843]
44. Donoho DL, Johnstone IM. Ideal spatial adaptation by wavelet shrinkage. Biometrika. Sep 1, 1994;81(3):425-455. [doi: 10.1093/biomet/81.3.425]
45. Nimptsch U. Disease-specific trends of comorbidity coding and implications for risk adjustment in hospital administrative data. Health Serv Res. Jun 2016;51(3):981-1001. [doi: 10.1111/1475-6773.12398] [Medline: 26741707]
46. Dunn OJ. Multiple comparisons among means. J Am Stat Assoc. Mar 1961;56(293):52-64. [doi: 10.1080/01621459.1961.10482090]
47. The R project for statistical computing. R Project. URL: https://www.R-project.org/ [Accessed 2026-08-17]
48. Sundararajan V, Quan H, Halfon P, et al. Cross-national comparative performance of three versions of the ICD-10 Charlson index. Med Care. Dec 2007;45(12):1210-1215. [doi: 10.1097/MLR.0b013e3181484347] [Medline: 18007172]
49. Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: a tutorial on multiple imputation. Can J Cardiol. Sep 2021;37(9):1322-1331. [doi: 10.1016/j.cjca.2020.11.010] [Medline: 33276049]
50. Loukides G, Liagouris J, Gkoulalas-Divanis A, Terrovitis M. Disassociation for electronic health record privacy. J Biomed Inform. Aug 2014;50:46-61. [doi: 10.1016/j.jbi.2014.05.009] [Medline: 24879898]
51. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care. Jan 1998;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
52. Suchikova Y, Tsybuliak N, Teixeira da Silva JA, Nazarovets S. GAIDeT (Generative AI Delegation Taxonomy): a taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. Account Res. Apr 2026;33(3):2544331. [doi: 10.1080/08989621.2025.2544331] [Medline: 40781729]

Abbreviations

CCI: Charlson Comorbidity Index
Destatis: German Federal Statistical Office
***ICD*:** *International Statistical Classification of Diseases*
***ICD-10-GM*:** *International Statistical Classification of Diseases, 10th Revision, German Modification*
***ICD-9-CM*:** *International Statistical Classification of Diseases, Ninth Revision, Clinical Modification*
KS: Kolmogorov-Smirnov
MAE: mean absolute error
MSE: mean squared error
NNLS: nonnegative least squares
RMSE: root mean square error
RQ: research question
UMM: University Hospital Mannheim

Edited by Andrew Coristine; peer-reviewed by Heber Anandan; submitted 14.Jul.2026; initial comments to author 12.Aug.2026; final revised version received 11.Oct.2026; accepted 19.Oct.2026; published 19.Nov.2026

Please cite as:
Kamdje Wabo G, Sokolowski PP, Jannesari Ladani M, Santhanam N, Hagmann M, Ganslandt T, Siegel F
Estimating the Charlson Comorbidity Index From Privacy-Truncated Diagnosis Codes: Single-Center Retrospective Cohort Study Evaluating Five Computational Strategies in 507,949 Inpatient Encounters
JMIR Med Inform 2026;14:e107041
URL: https://medinform.jmir.org/2026/1/e107041
doi: 10.2196/107041

© Gaetan Kamdje Wabo, Piotr Pawel Sokolowski, Mahboubeh Jannesari Ladani, Nandhini Santhanam, Michael Hagmann, Thomas Ganslandt, Fabian Siegel. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.