Online Resource 1 — Methods and parameters

Article: Synthetic tabular healthcare data in health informatics: fixed-size substitution and incremental-value controls for predictive utility

Journal: International Journal of Data Science and Analytics

Authors: Miguel Álvaro Andrade de Oliveira; Miguel Ricou; Alberto Freitas

Corresponding author: Miguel Álvaro Andrade de Oliveira; miguelalvaro13@gmail.com

Scope and provenance

Scientific methods, recorded parameter values and seed documentation for the supplied primary fixed-size benchmark. This is a non-code reading supplement, not an executable reproduction package. No new experiment, model fitting, split reconstruction or result recalculation was performed for this edition. Scientific primary tables and figures are in Online Resources 2 and 3; the additive control is in Online Resource 4.

Values below are taken from the v12 archive's resolved configurations and protocol manifests, not inferred solely from its default configuration. Exact source member paths, file SHA-256 values and the separately recorded config_sha256 values are in Recorded_configuration_provenance.json. A recorded configuration digest is reported as recorded and is not conflated with the byte hash of its JSON file.

Scientific question

Does replacing increasing fractions of an exactly fixed-size classifier-training set with synthetic observations preserve or improve predictive utility relative to training on all available real observations?

Fixed-size scenarios

For each dataset and replication, let N be the size of the original real training partition.

Scenario	Real rows	Synthetic rows	Classifier-training total
RR	N	0	N
RS25	N − round(0.25N)	round(0.25N)	N
RS50	N − round(0.50N)	round(0.50N)	N
RS75	N − round(0.75N)	round(0.75N)	N
SR	0	N	N

The synthetic generator is always fitted on the complete original real training partition. Only the composition of the downstream classifier-training frame changes. Validation and test partitions remain unchanged.

Real and synthetic memberships are selected without replacement using deterministic target-aware largest-remainder allocation and stable cryptographic seed derivation. Memberships are nested across proportions wherever class allocation permits. The same scenario frame is reused across all three downstream models.

Locked components

- Original train/validation/test partition
- Replication seed list
- Gaussian Copula, CTGAN and TVAE generators
- Logistic Regression, Random Forest and HistGradientBoostingClassifier
- Model hyperparameters
- Preprocessing and missing-value handling
- Classification threshold of 0.50
- Utility, fidelity, privacy-proxy, calibration and subgroup metrics
- Bootstrap mean confidence intervals (2,000 iterations; 95%)

- Ten replications

Statistical analysis

Primary paired inference is computed separately for each dataset × generator × downstream model × scenario × metric using the ten matched replications.

- Two-sided paired Wilcoxon signed-rank test (zero_method="wilcox")
- Paired rank-biserial correlation as the pairing-compatible effect size
- Cliff's Delta as a supplementary marginal/distributional effect size
- Holm correction within each dataset × downstream model × metric family across the twelve planned contrasts (three generators × four non-RR scenarios)

Because the train/validation/test partition is fixed, inferential results quantify stochastic pipeline variability conditional on that partition. They are not population-level sampling inference across alternative splits.

Interpretation boundary

The design controls the number of rows supplied to the downstream classifier. It does not equalize the generator's informational exposure: every generator is fitted to all N real training records, including records not directly included in the real component of RS25/RS50/RS75. Accordingly, the experiment estimates the effect of substituting direct real rows with synthetic representations learned from the complete training partition.

Split specification and limitations

The configured target proportions are 0.60/0.20/0.20, with stratification and group preservation enabled, and split/runtime seed 42. Diabetes 130-US groups encounters by patient_nbr; encounter_id and patient_nbr are excluded from predictors. Early Stage groups exact profiles of all 16 predictors, excluding the outcome; conflicting labels within a predictor profile are disallowed. CDC uses the stratified split rule without a patient-group field. A fixed seed alone does not establish identical split membership across grouped-split implementation versions.

The v12 audit identified version-dependent grouped splitting: archived Diabetes 130-US training size 60,822 and test prevalence 0.1146759691 matched the tested scikit-learn 1.3.2 combination (60,822/20,591/20,353), whereas the tested 1.8.0 combination yielded 61,060/20,352/20,354 and test prevalence 0.1116242508. Conversely, archived Early Stage sizes 312/104/104 matched the tested 1.8.0 combination; the tested 1.3.2 combination yielded 328/101/91. These diagnostics do not establish the exact historical software provenance of every output or identical membership merely from matching counts. No single universal software version has been validated to reconstruct all archived partitions; immutable split assignments are not supplied here. This limitation does not itself demonstrate data leakage or invalidate the stored numerical tables. Inference remains conditional on the archived fixed partitions.

Resolved parameters versus template defaults

The resolved primary configurations record ten replications and 2,000 bootstrap iterations, confidence level 0.95 and minimum subgroup size 30. The source configs/default.yaml instead contains a three-replication, 200-bootstrap template and lists Heart Failure; Heart Failure is not a primary dataset. The resolved configurations disable llm_tabular; no LLM generator is part of this benchmark. The three entries [101, 202, 303] remaining in the resolved replication_seeds field are not the complete ten-replication seed list: the protocol manifests record the complete list below. Unused mock-LLM settings are not experimental arms.

Exact model and active synthesis parameters

Component	Parameter	Recorded value
hist_gradient_boosting	enabled	true

hist_gradient_boosting	l2_regularization	0.0
hist_gradient_boosting	learning_rate	0.05
hist_gradient_boosting	max_depth	null
hist_gradient_boosting	max_iter	200
logistic_regression	C	1.0
logistic_regression	enabled	true
logistic_regression	max_iter	1000
logistic_regression	penalty	"l2"
logistic_regression	solver	"lbfgs"
random_forest	enabled	true
random_forest	max_depth	null
random_forest	min_samples_leaf	2
random_forest	min_samples_split	2
random_forest	n_estimators	300
Component	**Parameter**	**Recorded value**
ctgan	batch_size	500
ctgan	enabled	true
ctgan	epochs	100
ctgan	verbose	false
gaussian_copula	enabled	true
gaussian_copula	enforce_min_max_values	true
gaussian_copula	enforce_rounding	true
tvae	batch_size	500
tvae	enabled	true
tvae	epochs	100
tvae	verbose	false

Dataset configuration fields

cdc_diabetes_health_indicators

Field	Recorded value
enabled	true
include_in_primary_benchmark	true
leakage_prone_columns	["ID"]
subgroup_columns	["Sex", "Age"]
target_column_candidates	["Diabetes_binary", "diabetes_binary", "Diabetes_012"]

diabetes_130

Field	Recorded value
enabled	true
include_in_primary_benchmark	true
leakage_prone_columns	["encounter_id", "patient_nbr"]
positive_class	"<30"
split_group_column	"patient_nbr"
subgroup_columns	["gender", "race", "age"]
target_column	"readmitted"

early_stage_diabetes

Field	Recorded value
enabled	true
group_identical_records	true
include_in_primary_benchmark	true
leakage_prone_columns	[]
subgroup_columns	["Gender", "Age"]
target_column_candidates	["class", "Class"]

Recorded replication seeds

These index–seed pairs occur in the delivered primary protocol manifests for each of the three datasets. The incremental protocol specifies the same ten seeds; this is a protocol specification, not a claim that individual incremental configuration records are included.

Replication index	Seed
0	101

1	202
2	303
3	89250954
4	773956048
5	654571518
6	438878440
7	433015236
8	858597919
9	85945639

The S4.1 caption records percentile bootstrap resampling of ten replication-level means, 2,000 draws, seed 42. No historical counts of target-support fallback or repair events are asserted: the delivered event evidence is insufficient to establish them.

Source documents

- Supplementary_File_1.zip::Supplementary_File_1/EXPERIMENT_PROTOCOL.md — SHA-256 795fa4e2104e829e6d1a568bded6c14af62058446aa1015461136e9f4ef348cc.
- Supplementary_File_1.zip::Supplementary_File_1/configs/default.yaml — SHA-256 9bd1074367c4d4b9126a72922375178643e290c33b2b8aa3cc9592d944467328.
- Supplementary_File_1.zip::Supplementary_File_1/configs/datasets.yaml — SHA-256 22b401df7031547d7965b18e9a90623f1d4343337b786721c3a83ed4955ee70f.
- Supplementary_File_1.zip::Supplementary_File_1/configs/models.yaml — SHA-256 5ff9bb4a26bdbc12cff12869cf8e9bd4232db5130dfd648075ea0916affec87a.
- Supplementary_File_1.zip::Supplementary_File_1/configs/synthesis.yaml — SHA-256 0a931d475e5383ab15f621240d04882fc00acea15bcdfeddc4bcbee21dea3864.
- Supplementary_File_4.zip::Supplementary_File_4/protocol/INCREMENTAL_VALUE_CONTROL_PROTOCOL.md and figures/Supplementary_Figure_S4_1_Caption.md supply the incremental protocol and caption.

Recorded configuration provenance

The structured provenance record is represented below as indented JSON text. Source paths are historical identifiers; source files are not included.


```
{
 "scope": "Historical primary configuration references only; no raw observations or executable code.",
 "records": [
  {
   "config_sha256": "7b3dface9a1dc84bfeb03ce299ff365a63f07f24c9c97e551d10a09395d1abb1",
   "dataset": "cdc_diabetes_health_indicators",
   "experiment_namespace": "balanced_augmentation_benchmark",
   "replication_index": 0,
   "replication_seed": 101,
   "scenarios": [
    "RR",
    "RS25",
    "RS50",
    "RS75",
    "SR"
   ],
   "source_archive": "Supplementary_File_1.zip",
   "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/replication_00/summaries/protocol_manifest.json",
   "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/replication_00/summaries/resolved_config.json",
   "resolved_config_file_sha256":
"7b3dface9a1dc84bfeb03ce299ff365a63f07f24c9c97e551d10a09395d1abb1",
   "protocol_file_sha256": "d4530447c26d34c60aae637e8dac417bbb54dcb6fe69cf46d3e0cfaa125d41f6"
  },
  {
   "config_sha256": "b00a0cd91b52d4b182c134ce560f8898a5fbe943d0c0f84fd4b80f2d37d76b2c",
   "dataset": "cdc_diabetes_health_indicators",
   "experiment_namespace": "balanced_augmentation_benchmark",
   "replication_index": 1,
   "replication_seed": 202,
   "scenarios": [
    "RR",
    "RS25",
```

```
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_01/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_01/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"b00a0cd91b52d4b182c134ce560f8898a5fbe943d0c0f84fd4b80f2d37d76b2c",
    "protocol_file_sha256": "6cf519f333716b86c6ce4ce21b605c2cbcee9c280a65e4219baf67f0d788be6b"
  },
  {
    "config_sha256": "17b830cf747570abfb6df14af35c785772905fac735857966938af517d62a6ef",
    "dataset": "cdc_diabetes_health_indicators",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 2,
    "replication_seed": 303,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_02/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_02/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"17b830cf747570abfb6df14af35c785772905fac735857966938af517d62a6ef",
    "protocol_file_sha256": "fb9412ce1662ed55c7a60fa1a1b243c7d65688f1bfed52f38a5a228fc9b0a2f0"
  },
  {
```

```
    "config_sha256": "eff131e0556ffa08202d2c4a3daf33621e84d342e1560347f699ae58c30cea8d",
    "dataset": "cdc_diabetes_health_indicators",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 3,
    "replication_seed": 89250954,
    "scenarios": [
     "RR",
     "RS25",
     "RS50",
     "RS75",
     "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_03/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_03/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"eff131e0556ffa08202d2c4a3daf33621e84d342e1560347f699ae58c30cea8d",
    "protocol_file_sha256": "a35e4d3bd786e182b047d997a89d0ae2b8cfe55d68ab6329ce652115817cff9f"
   },
   {
    "config_sha256": "dac22fada3c19870a7e94d3f65b1f1616283f79826ae0fda6b2a4ba20472ed04",
    "dataset": "cdc_diabetes_health_indicators",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 4,
    "replication_seed": 773956048,
    "scenarios": [
     "RR",
     "RS25",
     "RS50",
     "RS75",
     "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
```

```
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/replication_04/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/replication_04/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"dac22fada3c19870a7e94d3f65b1f1616283f79826ae0fda6b2a4ba20472ed04",
    "protocol_file_sha256": "a44c79d6a03bc1a6cc816de9d7dd4b373f444e1fdf0df476f3b67d14f3972326"
  },
  {
    "config_sha256": "e8a0958efb05ff2d36f76cf152374f7b0d6aab1cf1e67836fabb1c22ca47532a",
    "dataset": "cdc_diabetes_health_indicators",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 5,
    "replication_seed": 654571518,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/replication_05/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/replication_05/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"e8a0958efb05ff2d36f76cf152374f7b0d6aab1cf1e67836fabb1c22ca47532a",
    "protocol_file_sha256": "af78b5fdd35385c96321be2d218b07ed378b0c7659a6082db4d74d4e270f71e4"
  },
  {
    "config_sha256": "5ee5eb9da339ece7647f6e937dd285cfb9e6cdee7df8c7d0048d2b28968399f7",
    "dataset": "cdc_diabetes_health_indicators",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 6,
    "replication_seed": 438878440,
```

```
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_06/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_06/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"5ee5eb9da339ece7647f6e937dd285cfb9e6cdee7df8c7d0048d2b28968399f7",
    "protocol_file_sha256": "56bbf59f53cbd9e4daead3ba2e7aaa5570ce3cebc2c2819e47be81683c8ff931"
  },
  {
    "config_sha256": "7ec70f5528a7f69f2a4859cf1c4b5f86308d1597541755f1b4e09bc76a7969a9",
    "dataset": "cdc_diabetes_health_indicators",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 7,
    "replication_seed": 433015236,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_07/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_07/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"7ec70f5528a7f69f2a4859cf1c4b5f86308d1597541755f1b4e09bc76a7969a9",
```

```
    "protocol_file_sha256": "2f5e98ae711477e23c2a67529b8f603a38bf14ca39eadb825f9265e650ac9adf"
   },
   {
    "config_sha256": "b8a3ab401e2ca2002c1b74c878693680bd247ac0c14f126325345dca4d29a5d2",
    "dataset": "cdc_diabetes_health_indicators",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 8,
    "replication_seed": 858597919,
    "scenarios": [
     "RR",
     "RS25",
     "RS50",
     "RS75",
     "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_08/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_08/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"b8a3ab401e2ca2002c1b74c878693680bd247ac0c14f126325345dca4d29a5d2",
    "protocol_file_sha256": "9e6d6425eb63c58d65d1160cf27ce3e5bdc5acfcac6ab9d28d53089df86c78ea"
   },
   {
    "config_sha256": "3d363b4d31ba28586d8b68bcfceff2b7c517c14619ca19a77c3a460b84223888",
    "dataset": "cdc_diabetes_health_indicators",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 9,
    "replication_seed": 85945639,
    "scenarios": [
     "RR",
     "RS25",
     "RS50",
     "RS75",
     "SR"
```

```
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_09/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/cdc_diabetes_health_indicators/r
eplication_09/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"3d363b4d31ba28586d8b68bcfceff2b7c517c14619ca19a77c3a460b84223888",
    "protocol_file_sha256": "904fb40d5db9e9cf6a9e46bf3775a29cedc3bc8d826b74d6c268078f0d4fadd1"
  },
  {
    "config_sha256": "809decb238b882e82bd24039fc57fe749fa27df9f0c5465b0f3d4d0f09630a5d",
    "dataset": "diabetes_130",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 0,
    "replication_seed": 101,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_00/sum
maries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_00/sum
maries/resolved_config.json",
    "resolved_config_file_sha256":
"809decb238b882e82bd24039fc57fe749fa27df9f0c5465b0f3d4d0f09630a5d",
    "protocol_file_sha256": "84203e8b5d3f6e3903d7b96ef49ab967a12cf14bc6f0f08a85e608eb04548397"
  },
  {
    "config_sha256": "452230cda0c8f4c179b003e2c62e9b98c4a9fe4f5c060444f3044fc043381757",
    "dataset": "diabetes_130",
    "experiment_namespace": "balanced_augmentation_benchmark",
```

```
    "replication_index": 1,
    "replication_seed": 202,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_01/sum
maries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_01/sum
maries/resolved_config.json",
    "resolved_config_file_sha256":
"452230cda0c8f4c179b003e2c62e9b98c4a9fe4f5c060444f3044fc043381757",
    "protocol_file_sha256": "c3db5d9201aaf6a857f84997cc4e85d48e49e0305efe465d366f3c1fecd2d651"
  },
  {
    "config_sha256": "83b943cc20b3cb3f8c16c778955eb8f6d22c7885969ccc3ee7f7c4a1f130fbda",
    "dataset": "diabetes_130",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 2,
    "replication_seed": 303,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_02/sum
maries/protocol_manifest.json",
```

```
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_02/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"83b943cc20b3cb3f8c16c778955eb8f6d22c7885969ccc3ee7f7c4a1f130fbda",
    "protocol_file_sha256": "2cc69b1eb9089d668fe20475e8c7cbeaa4666825bd7aaad9472cf7d224480d31"
   },
   {
    "config_sha256": "72340ece6c2944c21f1423e465a2f0f4fb3504b2ea526bb997c72d7cc8bb555a",
    "dataset": "diabetes_130",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 3,
    "replication_seed": 89250954,
    "scenarios": [
     "RR",
     "RS25",
     "RS50",
     "RS75",
     "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_03/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_03/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"72340ece6c2944c21f1423e465a2f0f4fb3504b2ea526bb997c72d7cc8bb555a",
    "protocol_file_sha256": "e176204c17b8293a8e79158f1e7bea60e8c716dafe281047c3f33c4a8d9e5447"
   },
   {
    "config_sha256": "a6e9e5ae276bf7aad36ab4af6be24b5c8c4bb5bcace488e5dd56b554c7d966a5",
    "dataset": "diabetes_130",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 4,
    "replication_seed": 773956048,
    "scenarios": [
     "RR",
```

```
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_04/sum
maries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_04/sum
maries/resolved_config.json",
    "resolved_config_file_sha256":
"a6e9e5ae276bf7aad36ab4af6be24b5c8c4bb5bcace488e5dd56b554c7d966a5",
    "protocol_file_sha256": "284cb2950ca192891881df458b2cc655a03b303007fc3f8c332d5ee326bdca8e"
  },
  {
    "config_sha256": "9b70ad49c57498812c84d8016c94e01950095d23dd5e5a4ed648e891a9572381",
    "dataset": "diabetes_130",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 5,
    "replication_seed": 654571518,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_05/sum
maries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_05/sum
maries/resolved_config.json",
    "resolved_config_file_sha256":
"9b70ad49c57498812c84d8016c94e01950095d23dd5e5a4ed648e891a9572381",
    "protocol_file_sha256": "96289359984174c737c8261fd8af5445146a9f5f1146b0cb0bdcb86f1d3bd9e5"
  },
```

```
  {
    "config_sha256": "9280e528ffa97ab71ee8f9a300699b2d1ad743af36141ec565ddbf68851787f3",
    "dataset": "diabetes_130",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 6,
    "replication_seed": 438878440,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_06/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_06/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"9280e528ffa97ab71ee8f9a300699b2d1ad743af36141ec565ddbf68851787f3",
    "protocol_file_sha256": "d9b8a2aaf7d503cd0284ef8f713cbeb3ee4edc2ec0f78a4dfd1ee177edbbc7ad"
  },
  {
    "config_sha256": "9108d5cd4804dfaf6175933d9651b5d707f35cd8c4ce96f36b19b341d23388d8",
    "dataset": "diabetes_130",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 7,
    "replication_seed": 433015236,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
```

```
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_07/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_07/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"9108d5cd4804dfaf6175933d9651b5d707f35cd8c4ce96f36b19b341d23388d8",
    "protocol_file_sha256": "7fd89c32cec84b2280460ba3843cf13b9b5f6cc0883ca49ee65af2c1b5f2eeed"
   },
   {
    "config_sha256": "684ab117a779b94bc2721aa39577b4e513ae4c7e395ea149a85bcd080f895ae3",
    "dataset": "diabetes_130",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 8,
    "replication_seed": 858597919,
    "scenarios": [
     "RR",
     "RS25",
     "RS50",
     "RS75",
     "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_08/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_08/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"684ab117a779b94bc2721aa39577b4e513ae4c7e395ea149a85bcd080f895ae3",
    "protocol_file_sha256": "4a7eba91d5e5eb2b7e93fdb8068815745c38404c11e88d2cec363f83fc7362b3"
   },
   {
    "config_sha256": "8bab40fbe51819ea5e999ce3b78fdf69087e0b91ec6e1f91b61382f37e41668d",
    "dataset": "diabetes_130",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 9,
    "replication_seed": 85945639,
```

```
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_09/sum
maries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/diabetes_130/replication_09/sum
maries/resolved_config.json",
    "resolved_config_file_sha256":
"8bab40fbe51819ea5e999ce3b78fdf69087e0b91ec6e1f91b61382f37e41668d",
    "protocol_file_sha256": "fe0cc2e06d55832db941cd4c3e3e3eb83b5b072ded6f60fc6859e91d074b07cc"
  },
  {
    "config_sha256": "34f69514d1861303c10600f5c5d59c15adfc5d8e64088f044d9b334e3b35f0a2",
    "dataset": "early_stage_diabetes",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 0,
    "replication_seed": 101,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
00/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
00/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"34f69514d1861303c10600f5c5d59c15adfc5d8e64088f044d9b334e3b35f0a2",
```

```
    "protocol_file_sha256": "a9e8319a8b9bdd4b2a3e793e49bcb8f6a26aa22acb86af164fae248ac4cc20f9"
   },
   {
    "config_sha256": "6acb49c65bb1cfd4548b7e5d7c75d07b8b9e3d2b4dd131bc4e2816cb01c7bc5e",
    "dataset": "early_stage_diabetes",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 1,
    "replication_seed": 202,
    "scenarios": [
     "RR",
     "RS25",
     "RS50",
     "RS75",
     "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
01/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
01/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"6acb49c65bb1cfd4548b7e5d7c75d07b8b9e3d2b4dd131bc4e2816cb01c7bc5e",
    "protocol_file_sha256": "24fda4f40b12e8153e987494abd56d81846716394fd8a6516957c105ddaa5c0e"
   },
   {
    "config_sha256": "662f1b510f29084c26c057c3d79748b5e87b29ce96275471972f13d824f6a582",
    "dataset": "early_stage_diabetes",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 2,
    "replication_seed": 303,
    "scenarios": [
     "RR",
     "RS25",
     "RS50",
     "RS75",
     "SR"
```

```
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
02/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
02/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"662f1b510f29084c26c057c3d79748b5e87b29ce96275471972f13d824f6a582",
    "protocol_file_sha256": "67b7f49147a2240f06880f81de39ffc1a6039b9e64d7489bc6be060b46f3b89f"
  },
  {
    "config_sha256": "66613e96a0cb3ba4b3ffc7e1b0d638ff133277e2b5cc6c2cc950668a39059e0b",
    "dataset": "early_stage_diabetes",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 3,
    "replication_seed": 89250954,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
03/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
03/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"66613e96a0cb3ba4b3ffc7e1b0d638ff133277e2b5cc6c2cc950668a39059e0b",
    "protocol_file_sha256": "b4490581eaa5496db53c406d6612c128669ace8259952ad961d44ca6eaf04183"
  },
  {
    "config_sha256": "1a16e9b8c0ff0ce76bf1f3d28b5bf7321c49667c330c133df3a8cfc880294c83",
    "dataset": "early_stage_diabetes",
    "experiment_namespace": "balanced_augmentation_benchmark",
```

```
    "replication_index": 4,
    "replication_seed": 773956048,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
04/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
04/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"1a16e9b8c0ff0ce76bf1f3d28b5bf7321c49667c330c133df3a8cfc880294c83",
    "protocol_file_sha256": "107b8d103836692c64b83773297938fac9f64bb633691058437761be629c9943"
  },
  {
    "config_sha256": "be8accf6c62129aec43642b1da1c8e4e2a6ed13a0fcfadf0ef44afa481773c92",
    "dataset": "early_stage_diabetes",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 5,
    "replication_seed": 654571518,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
05/summaries/protocol_manifest.json",
```

```
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
05/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"be8accf6c62129aec43642b1da1c8e4e2a6ed13a0fcfadf0ef44afa481773c92",
    "protocol_file_sha256": "79ee8b6f0c89cd46461a8f423b866cff41fe7d79130ea834f909a081d0a0dd0d"
   },
   {
    "config_sha256": "6f815a0b33d6cf728ab5276c979033afa56c7ea502af0abb2c39915a0f49a649",
    "dataset": "early_stage_diabetes",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 6,
    "replication_seed": 438878440,
    "scenarios": [
     "RR",
     "RS25",
     "RS50",
     "RS75",
     "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
06/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
06/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"6f815a0b33d6cf728ab5276c979033afa56c7ea502af0abb2c39915a0f49a649",
    "protocol_file_sha256": "84fa8f05f1657c8c0c28872e23cf05a91eb35dd2d6eea2015f7e474c82aecf00"
   },
   {
    "config_sha256": "45e3ef936f67100b668166654feeed2301180b936a13831dc909eba5e635111f",
    "dataset": "early_stage_diabetes",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 7,
    "replication_seed": 433015236,
    "scenarios": [
     "RR",
```

```
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
07/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
07/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"45e3ef936f67100b668166654feeed2301180b936a13831dc909eba5e635111f",
    "protocol_file_sha256": "3f1027ab767434b5c2ca999487f394c1867e0575add340cc5ce3e074b6f81d8b"
  },
  {
    "config_sha256": "54d45100d92f00d429be0612ad99f068efa9484b9414a5155f67746b7e3929bb",
    "dataset": "early_stage_diabetes",
    "experiment_namespace": "balanced_augmentation_benchmark",
    "replication_index": 8,
    "replication_seed": 858597919,
    "scenarios": [
      "RR",
      "RS25",
      "RS50",
      "RS75",
      "SR"
    ],
    "source_archive": "Supplementary_File_1.zip",
    "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
08/summaries/protocol_manifest.json",
    "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
08/summaries/resolved_config.json",
    "resolved_config_file_sha256":
"54d45100d92f00d429be0612ad99f068efa9484b9414a5155f67746b7e3929bb",
    "protocol_file_sha256": "fd1d59fe90f00efcb1bf0b26956dfe2d990af641cb7e6dd47d9e150fb661a20e"
  },
```

```
    {
      "config_sha256": "b35fba791a527e8a754e47b1e2fde9f137e7d89980e0bd2bafa1d9a7d4ccf8f5",
      "dataset": "early_stage_diabetes",
      "experiment_namespace": "balanced_augmentation_benchmark",
      "replication_index": 9,
      "replication_seed": 85945639,
      "scenarios": [
        "RR",
        "RS25",
        "RS50",
        "RS75",
        "SR"
      ],
      "source_archive": "Supplementary_File_1.zip",
      "protocol_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
09/summaries/protocol_manifest.json",
      "resolved_config_member":
"Supplementary_File_1/outputs/balanced_augmentation_benchmark/slices/early_stage_diabetes/replication_
09/summaries/resolved_config.json",
      "resolved_config_file_sha256":
"b35fba791a527e8a754e47b1e2fde9f137e7d89980e0bd2bafa1d9a7d4ccf8f5",
      "protocol_file_sha256": "47c1bbb22b03085d780dbebbbca4e07d7142ef60afcd18d89747686736f82989"
    }
  ],
  "seed_index_mapping": {
    "cdc_diabetes_health_indicators": [
      [
        0,
        101
      ],
      [
        1,
        202
      ],
      [
        2,
        303
```

```
  ],
  [
    3,
    89250954
  ],
  [
    4,
    773956048
  ],
  [
    5,
    654571518
  ],
  [
    6,
    438878440
  ],
  [
    7,
    433015236
  ],
  [
    8,
    858597919
  ],
  [
    9,
    85945639
  ]
 ],
 "diabetes_130": [
  [
    0,
    101
  ],
  [
    1,
```

```
    202
   ],
   [
    2,
    303
   ],
   [
    3,
    89250954
   ],
   [
    4,
    773956048
   ],
   [
    5,
    654571518
   ],
   [
    6,
    438878440
   ],
   [
    7,
    433015236
   ],
   [
    8,
    858597919
   ],
   [
    9,
    85945639
   ]
  ],
  "early_stage_diabetes": [
   [
```

```
  0,
  101
],
[
  1,
  202
],
[
  2,
  303
],
[
  3,
  89250954
],
[
  4,
  773956048
],
[
  5,
  654571518
],
[
  6,
  438878440
],
[
  7,
  433015236
],
[
  8,
  858597919
],
[
  9,
```

```
        85945639
      ]
    ]
  }
}
```