Supplementary Material

for

Nationwide prediction of type 2 diabetes comorbidities

Piotr Dworzynski\textsuperscript{1,2}, Martin Aasbrenn\textsuperscript{1,3}, Klaus Rostgaard\textsuperscript{2}, Mads Melbye\textsuperscript{2,4,5}, Thomas Alexander Gerds\textsuperscript{6}, Henrik Hjalgrim\textsuperscript{2,5}, Tune H Pers\textsuperscript{1,2,4}\textsuperscript{+}

\textsuperscript{1}The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
\textsuperscript{2}Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark
\textsuperscript{3}Department of Geriatrics and Internal Medicine, Bispebjerg and Frederiksberg Hospital, Copenhagen
\textsuperscript{4}Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark
\textsuperscript{5}Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA
\textsuperscript{6}Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

\textsuperscript{+}These authors contributed equally to this work

Correspondence to: tune.pers@sund.ku.dk

This file includes:

- Supplementary Note 1: Death as a competing risk
- Supplementary Figures 1 to 9
- Supplementary Tables 1 to 10
Supplementary Note 1: Death as a competing risk

Individuals who died during the prediction horizon before potentially being diagnosed with the comorbidity were categorized as non-cases. Consequently, death constituted a competing risk to the comorbidity diagnosis. To investigate the extent to which death censoring could impact predictions of the baseline and gradient boosting models we conducted two analyses. First, for each comorbidity and the gradient boosting and the reference model, we investigated five-year incidence of all-cause mortality in each population percentile (in the distribution derived by ranking individuals based on their predicted risk of a given comorbidity). We observed that for all comorbidities incidence of all-cause mortality was generally increasing with the predicted risk of the comorbidity (see Supplementary Fig. 8b for the results for CKD and Supplementary Fig. 9 for the other four comorbidities). Second, we trained and evaluated prediction performance for all-cause mortality using the same procedure as for T2D comorbidities. We found that gradient boosting significantly outperformed all other models (Supplementary Table 10) with AUROCs substantially higher compared to the best AUROCs observed for the T2D comorbidities (AUROC=0.87 vs. HF AUROC=0.80). Furthermore, the gradient boosting feature importances for all-cause mortality, in contrast to the comorbidity prediction feature importances, suggested a more prominent role of hospital diagnoses (among top 7 features three were diagnoses of malignant neoplasms) compared to canonical features and prescriptions (cumulative importances of 32%, 19% and 29%, compared to average cumulative importances of 24.2%, 22.2% and 36.8%, respectively; Supplementary Fig. 8c). These results show that competing risk of death is especially prominent among individuals predicted to be at highest comorbidity risk and that Danish health registers contain information highly predictive of all-cause mortality.

The register-based ML models, due to their high flexibility, could potentially better account for this risk as firstly, the health registers may contain variables, such as diagnoses of terminal diseases e.g. cancers, highly predictive of death. Secondly, more flexible models such as random forest or gradient boosting can better leverage non linear relationships between the features and the outcome by accounting that e.g. very advanced age, otherwise typically positively correlated with a comorbidity risk, may decrease overall comorbidity risk due to increased risk of mortality. In the analogous task of five-year all-cause mortality prediction at first T2D diagnosis, gradient boosting significantly outperformed the reference model achieving a relatively high AUROC of 0.87 when identifying subgroups at high (95th percentile risk ratio of 4.53) risk of death. Furthermore, the best all-cause mortality prediction model was able to identify relatively rare but highly predictive hospital diagnosis features such as cancer diagnoses. Lastly, for HF, CVD, and CKD individuals predicted to be at highest risk of developing a given comorbidity by a register-based model had a lower incidence of all-cause mortality than their counterparts selected by the reference model. This indicates that ML register-based models can outperform the reference model through better estimation of individuals risk of death and weighing it against the risk of a given comorbidity.
Supplementary Figure 1. Histogram of lengths of hospital admissions during which individuals received their first hospital diagnosis of T2D. Individuals with a prior T2D prescription-based diagnosis were excluded. The hospital admissions were limited to those lasting less than 100 days. T2D, type 2 diabetes.

Supplementary Table 1. Overview of study population characteristics among all newly diagnosed type 2 diabetics (T2D population) and two chronic kidney disease comorbidity populations with buffer period set to 30 and 60 days respectively. RFV, register feature vector; #, number of; BPL, buffer period length.
### Chronic kidney disease (BPL 30 days) (incidence: 0.03)

| Model                          | AUROC          | ΔAUROC<sub>RLR</sub> | ΔAUROC<sub>LRL</sub> | ΔAUROC<sub>RF</sub> |
|-------------------------------|----------------|----------------------|----------------------|---------------------|
| Reference, logistic regression (RLR) | 0.71 (0.69 — 0.73) |                       |                      |                     |
| Logistic regression (LR)       | 0.74 (0.72 — 0.76) | 0.04 (0.02 — 0.05)  |                      |                     |
| Random forest (RF)             | 0.74 (0.72 — 0.76) | 0.03 (0.01 — 0.05) | 0.00 (-0.02 — 0.01) |                     |
| Gradient boosting (GB)         | 0.77 (0.76 — 0.79) | 0.07 (0.05 — 0.08) | 0.03 (0.02 — 0.04) | 0.04 (0.02 — 0.05) |

### Chronic kidney disease (BPL 60 days) (incidence: 0.03)

| Model                          | AUROC          | ΔAUROC<sub>RLR</sub> | ΔAUROC<sub>LRL</sub> | ΔAUROC<sub>RF</sub> |
|-------------------------------|----------------|----------------------|----------------------|---------------------|
| Reference, logistic regression (RLR) | 0.70 (0.68 — 0.72) |                       |                      |                     |
| Logistic regression (LR)       | 0.74 (0.72 — 0.76) | 0.04 (0.03 — 0.06)  |                      |                     |
| Random forest (RF)             | 0.73 (0.71 — 0.75) | 0.03 (0.01 — 0.05) | -0.01 (-0.03 — 0.01) |                     |
| Gradient boosting (GB)         | 0.77 (0.75 — 0.78) | 0.07 (0.05 — 0.08) | 0.03 (0.02 — 0.04) | 0.04 (0.02 — 0.05) |

**Supplementary Table 2.** Comparison of AUROC measures for each prediction models best parameterization between chronic kidney disease comorbidity populations with buffer period set to 30 and 60 days respectively. We applied a reference- and three register-based models on fifteen years of health register data comprising hospital diagnoses, hospital procedures, drug prescriptions and interactions with primary care contractors to predict five-year risk for chronic kidney disease comorbidity. For each comorbidity, prediction was performed on a T2D population free of that comorbidity at the date of prediction (date of individuals first T2D diagnosis). The reference model was a logistic ridge regression based on canonical features: age, sex, country or region of birth and date of first T2D diagnosis as well as their interactions, while the register-based models were logistic ridge regression, random forest and gradient boosting based on the canonical features as well as hospital diagnoses, hospital procedures, drug prescriptions and interactions with primary care extracted from Danish health registers. Incidences are proportions of cases within comorbidities sub-population at the end of the prediction horizon. Value ranges in brackets represent 95% confidence intervals based on bootstrap sampling. AUROC, area under receiver operating characteristic curve.
|                      | T2D population |   | MI     |   | T2D population |   | MI     |   |
|----------------------|----------------|---|--------|---|----------------|---|--------|---|
| # Individuals        | 181,100        |   | 169,729|   | 169,972        |   | 169,729|   |
| # Cases              | 8,222          |   | 6,039  |   | 6,039          |   | 6,039  |   |
| # Non-cases          | 161,507        |   | 163,933|   | 163,933        |   | 163,933|   |
| % Women              | 46.98          |   | 48.13  |   | 48.13          |   | 48.13  |   |
| Median age at T2D diagnosis | 61.71         |   | 60.95  |   | 61.24          |   | 61.24  |   |
| # Days until outcome | 1672.9          |   | 1546.60|   | 1546.60        |   | 1546.60|   |
| # Features in RFV    | 6,093          |   | 6,093  |   | 6,093          |   | 6,093  |   |

|                      | ST             |   | CYD    |   | CKD            |   | CYD    |   |
|----------------------|----------------|---|--------|---|----------------|---|--------|---|
| # Individuals        | 169,870        |   | 103,971|   | 178,586        |   | 103,971|   |
| # Cases              | 7,467          |   | 28,638 |   | 173,225        |   | 7,323  |   |
| # Non-cases          | 162,403        |   | 75,333 |   | 72,466         |   | 147,451|   |
| % Women              | 47.37          |   | 49.99  |   | 47.13          |   | 47.13  |   |
| Median age at T2D diagnosis | 61.07         |   | 57.53  |   | 61.61          |   | 58.17  |   |
| # Days until outcome | 1636.50        |   | 1465.00|   | 2060.20        |   | 1452.00|   |
| # Features in RFV    | 6,093          |   | 6,093  |   | 6,093          |   | 6,093  |   |

**Supplementary Table 3.** Population characteristics at different dates of prediction. T2D, type 2 diabetes; HF, heart failure; MI, myocardial infarction; ST, stroke; CVD, cardiovascular disease; CKD, chronic kidney disease; RFV, register feature vector; #, number of.
**Logistic regression (LR)**

- penalty: L2
- solver: lbfgs
- C: 0.6, 0.7, 0.8
- max_iter: 300, 400, 500

**Random forest (RF)**

- max_features: sqrt
- criterion: gini
- n_estimators: 1000, 1500
- max_depth: 12, 14, 16

**Gradient boosting (GB)**

- n_estimators: 200
- booster: gbtree
- eval_metric: auc
- learning_rate: 0.025, 0.05, 0.1
- max_depth: 2, 3, 4, 5

**Supplementary Table 4.** Overview of model hyperparameters that were evaluated. For each model type, all combinations of listed hyperparameters were tested to identify those that led to the best average AUROC using 3-fold cross validation on the training set. Parameter names listed correspond to naming in software implementation (Python modules scikit-learn for LR and RF and xgboost for GB). Due to class imbalance (difference between number of cases and non-cases in each population) training error for class representing cases was scaled up by the inverse proportion between cases and non-cases.
Supplementary Figure 2. Validation set calibration curves of uncalibrated best models for each comorbidity and all-cause mortality with date of prediction set to an individual’s first type 2 diabetes diagnosis.
Supplementary Figure 3. Validation set calibration error (average difference between observed and predicted outcome probabilities for each predicted probability percentile) of uncalibrated best models for each comorbidity and all-cause mortality with date of prediction set to an individual’s first type 2 diabetes diagnosis.
Supplementary Figure 4. Validation set calibration curves of calibrated best models for each comorbidity and all-cause mortality with date of prediction set to an individual’s first type 2 diabetes diagnosis. Each best model was calibrated using Platt scaling method on the test set.
Supplementary Figure 5. Validation set calibration error (average difference between observed and predicted outcome probabilities for each predicted probability percentile) of calibrated best models for each comorbidity and all-cause mortality with date of prediction set to an individual’s first type 2 diabetes diagnosis. Each best model was calibrated using Platt scaling method on the test set.
Supplementary Figure 6. (a) Five-year incidence of hospital diagnosis of stroke for population percentiles ranked by risk as predicted by the best gradient boosting (blue) and the best baseline (orange) models. (b) Individuals were ranked according to their predicted risk of stroke by the best gradient boosting (blue) and the best baseline (orange) models. For a number of thresholds, shown are risk ratios, calculated as the stroke incidence of individuals ranking above that thresholds over ST incidence in the entire study population. 95% confidence interval (shaded areas) were obtained through bootstrap sampling. (c) 50 most predictive features for stroke according to the best gradient boosting models feature importances.
Table 5. Area under receiver operating characteristic curve (AUROC) performance for each tested parametrization for heart failure, myocardial infarction, stroke and cardiovascular disease. Date of prediction was set to date of first type 2 diabetes diagnosis. Hyperparameter names correspond to parameter names from the python sklearn package. GB, gradient boosting; RF, random forest; LR, register-based logistic regression; RLR, reference logistic regression.
| Model | Chronic kidney disease | AUROC | Model | All-cause mortality | AUROC |
|-------|------------------------|-------|-------|---------------------|-------|
| GB max_depth: 2 learning_rate: 0.1 n_estimators: 200 | 0.754 | GB max_depth: 4 learning_rate: 0.1 n_estimators: 200 | 0.864 |
| GB max_depth: 3 learning_rate: 0.05 n_estimators: 200 | 0.754 | GB max_depth: 5 learning_rate: 0.1 n_estimators: 200 | 0.863 |
| GB max_depth: 3 learning_rate: 0.1 n_estimators: 200 | 0.753 | GB max_depth: 3 learning_rate: 0.1 n_estimators: 200 | 0.863 |
| GB max_depth: 3 learning_rate: 0.025 n_estimators: 200 | 0.752 | GB max_depth: 4 learning_rate: 0.05 n_estimators: 200 | 0.862 |
| GB max_depth: 4 learning_rate: 0.01 n_estimators: 200 | 0.752 | GB max_depth: 2 learning_rate: 0.1 n_estimators: 200 | 0.861 |
| GB max_depth: 2 learning_rate: 0.025 n_estimators: 200 | 0.751 | GB max_depth: 3 learning_rate: 0.05 n_estimators: 200 | 0.860 |
| GB max_depth: 5 learning_rate: 0.1 n_estimators: 200 | 0.749 | GB max_depth: 4 learning_rate: 0.025 n_estimators: 200 | 0.859 |
| GB max_depth: 3 learning_rate: 0.025 n_estimators: 200 | 0.749 | GB max_depth: 4 learning_rate: 0.025 n_estimators: 200 | 0.857 |
| GB max_depth: 5 learning_rate: 0.05 n_estimators: 200 | 0.747 | GB max_depth: 2 learning_rate: 0.05 n_estimators: 200 | 0.856 |
| GB max_depth: 3 learning_rate: 0.1 n_estimators: 200 | 0.745 | GB max_depth: 3 learning_rate: 0.025 n_estimators: 200 | 0.854 |
| LR C: 0.6 max_iter: 500 | 0.730 | GB max_depth: 2 learning_rate: 0.025 n_estimators: 200 | 0.847 |
| LR C: 0.6 max_iter: 400 | 0.729 | RF n_estimators: 1500 max_depth: 16 | 0.845 |
| RF n_estimators: 1500 max_depth: 12 | 0.729 | RF n_estimators: 1000 max_depth: 16 | 0.848 |
| LR C: 0.8 max_iter: 500 | 0.729 | RF n_estimators: 1500 max_depth: 14 | 0.842 |
| LR C: 0.8 max_iter: 400 | 0.729 | RF n_estimators: 1000 max_depth: 14 | 0.842 |
| RF n_estimators: 1000 max_depth: 12 | 0.729 | RF n_estimators: 1500 max_depth: 12 | 0.838 |
| LR C: 0.7 max_iter: 500 | 0.727 | LR C: 0.8 max_iter: 500 | 0.827 |
| LR C: 0.7 max_iter: 400 | 0.727 | LR C: 0.7 max_iter: 500 | 0.827 |
| LR C: 0.7 max_iter: 300 | 0.726 | LR C: 0.6 max_iter: 500 | 0.826 |
| LR C: 0.7 max_iter: 300 | 0.726 | LR C: 0.7 max_iter: 500 | 0.826 |
| RF n_estimators: 1500 max_depth: 14 | 0.718 | LR C: 0.6 max_iter: 400 | 0.823 |
| RF n_estimators: 1000 max_depth: 14 | 0.717 | LR C: 0.6 max_iter: 500 | 0.823 |
| RF n_estimators: 1500 max_depth: 16 | 0.701 | LR C: 0.6 max_iter: 300 | 0.822 |
| RF n_estimators: 1000 max_depth: 16 | 0.700 | LR C: 0.7 max_iter: 300 | 0.822 |
| RLR C: 0.6 max_iter: 500 | 0.693 | RLR C: 0.6 max_iter: 300 | 0.796 |
| RLR C: 0.6 max_iter: 400 | 0.693 | RLR C: 0.6 max_iter: 500 | 0.796 |
| RLR C: 0.6 max_iter: 300 | 0.693 | RLR C: 0.6 max_iter: 500 | 0.796 |
| RLR C: 0.7 max_iter: 500 | 0.693 | RLR C: 0.7 max_iter: 500 | 0.796 |
| RLR C: 0.7 max_iter: 400 | 0.693 | RLR C: 0.7 max_iter: 500 | 0.796 |
| RLR C: 0.7 max_iter: 300 | 0.693 | RLR C: 0.8 max_iter: 300 | 0.796 |
| RLR C: 0.8 max_iter: 400 | 0.693 | RLR C: 0.8 max_iter: 400 | 0.796 |
| RLR C: 0.8 max_iter: 500 | 0.693 | RLR C: 0.8 max_iter: 500 | 0.796 |

Supplementary Table 6. Area under receiver operating characteristic curve (AUROC) performance for each tested parametrization for chronic kidney disease and all cause mortality. Date of prediction was set to date of first type 2 diabetes diagnosis. Hyperparameter names correspond to parameter names from the the python sklearn package. GB, gradient boosting; RF, random forest; RLR, register-based logistic regression; LR, reference logistic regression.
Supplementary Figure 7. Top seven most predictive gradient boosting features. For each comorbidity and ACM shown are the top seven features according to gradient boosting feature importance. Feature importance is an estimate of feature’s relative contribution to outcome prediction. Box plots show a distribution of a given continuous feature (e.g. age, an interaction between age and sex) among cases and non-cases within validation set. Box plot whiskers represent lowest and highest observations still within 1.5 inter quantile range. To comply with Danish data protection rules, these values as well as values representing 25th, 50th and 75th percentiles were obtained by averaging five closest observations. Bar plots, describing count based features (e.g. count of a given drug prescription, count of a given diagnosis), show the proportion of validation set cases and non-cases with at least a single observation of that feature. HF, heart failure; MI, myocardial infarction; ST, stroke; CVD, cardiovascular disease; CKD, chronic kidney disease; ACM, all-cause mortality; D, diagnosis of; P, prescription of; MO, modulators of; STE, st elevation; RA, renin-angiotensin; ODGRPIS, other disorders of glucose regulation and pancreatic internal secretion; RIPLMD, hmg coa reductase inhibitors and plain lipid modifying drugs.
| Event                  | Time (tp) | Gradient Boosting AUROC | Reference AUROC |
|-----------------------|-----------|-------------------------|-----------------|
| **Heart failure**     | 0         | 0.80 (0.78 — 0.81)      | 0.74 (0.72 — 0.75) |
|                       | 1         | 0.80 (0.78 — 0.81)      | 0.74 (0.72 — 0.76) |
|                       | 2         | 0.80 (0.78 — 0.81)      | 0.74 (0.72 — 0.76) |
|                       | 3         | 0.81 (0.79 — 0.83)      | 0.75 (0.73 — 0.77) |
|                       | 4         | 0.81 (0.79 — 0.82)      | 0.74 (0.72 — 0.75) |
| **Myocardial infarction** | 0       | 0.71 (0.69 — 0.73)      | 0.68 (0.65 — 0.70) |
|                       | 1         | 0.73 (0.71 — 0.75)      | 0.68 (0.66 — 0.71) |
|                       | 2         | 0.74 (0.71 — 0.76)      | 0.69 (0.67 — 0.71) |
|                       | 3         | 0.72 (0.70 — 0.75)      | 0.69 (0.67 — 0.72) |
|                       | 4         | 0.73 (0.70 — 0.75)      | 0.68 (0.66 — 0.71) |
| **Stroke**            | 0         | 0.72 (0.70 — 0.74)      | 0.71 (0.69 — 0.73) |
|                       | 1         | 0.75 (0.73 — 0.77)      | 0.70 (0.68 — 0.72) |
|                       | 2         | 0.75 (0.73 — 0.77)      | 0.71 (0.69 — 0.73) |
|                       | 3         | 0.75 (0.73 — 0.77)      | 0.72 (0.70 — 0.74) |
|                       | 4         | 0.75 (0.73 — 0.77)      | 0.72 (0.70 — 0.74) |
| **Cardiovascular disease** | 0     | 0.69 (0.68 — 0.70)      | 0.66 (0.64 — 0.67) |
|                       | 1         | 0.71 (0.70 — 0.73)      | 0.67 (0.66 — 0.68) |
|                       | 2         | 0.72 (0.71 — 0.74)      | 0.68 (0.67 — 0.70) |
|                       | 3         | 0.73 (0.72 — 0.74)      | 0.68 (0.67 — 0.69) |
|                       | 4         | 0.73 (0.72 — 0.74)      | 0.69 (0.67 — 0.70) |
| **Chronic kidney disease** | 0     | 0.77 (0.76 — 0.79)      | 0.71 (0.69 — 0.73) |
|                       | 1         | 0.77 (0.75 — 0.78)      | 0.68 (0.66 — 0.70) |
|                       | 2         | 0.77 (0.75 — 0.79)      | 0.70 (0.68 — 0.72) |
|                       | 3         | 0.79 (0.77 — 0.80)      | 0.72 (0.70 — 0.74) |
|                       | 4         | 0.79 (0.77 — 0.81)      | 0.70 (0.68 — 0.72) |
| **All-cause mortality** | 0     | 0.87 (0.86 — 0.87)      | 0.80 (0.79 — 0.81) |
|                       | 1         | 0.87 (0.87 — 0.88)      | 0.81 (0.80 — 0.82) |
|                       | 2         | 0.88 (0.87 — 0.88)      | 0.81 (0.80 — 0.82) |
|                       | 3         | 0.88 (0.87 — 0.88)      | 0.82 (0.81 — 0.83) |
|                       | 4         | 0.87 (0.87 — 0.88)      | 0.80 (0.79 — 0.81) |

**Supplementary Table 7.** Prediction performance at one, two, three and four years after individual’s first type 2 diabetes diagnosis. \( tp \), time of prediction counted in years since individual’s first type 2 diabetes diagnosis; \( t \), outcome incidence within study population; AUROC, area under receiver operating characteristic curve which confidence intervals were obtained by bootstrap sampling procedure.
| Event                          | Reference, logistic regression | Gradient boosting |
|-------------------------------|-------------------------------|-------------------|
| Heart failure (incidence: 0.04) | AUROC                         | AUROC             |
|                               | **Reference, logistic regression** | **0.74 (0.72 — 0.76)** | **0.80 (0.78 — 0.81)** |
|                               | **Gradient boosting**          | **0.73 (0.71 — 0.75)** | **0.80 (0.78 — 0.81)** |
| Myocardial infarction (incidence: 0.03) | AUROC                         | AUROC             |
|                               | **Reference, logistic regression** | **0.68 (0.66 — 0.71)** | **0.74 (0.71 — 0.76)** |
|                               | **Gradient boosting**          | **0.73 (0.71 — 0.77)** | **0.74 (0.71 — 0.76)** |
| Stroke (incidence: 0.03)      | AUROC                         | AUROC             |
|                               | **Reference, logistic regression** | **0.71 (0.69 — 0.73)** | **0.74 (0.71 — 0.77)** |
|                               | **Gradient boosting**          | **0.71 (0.70 — 0.73)** | **0.75 (0.73 — 0.77)** |
| Cardiovascular disease (incidence: 0.24) | AUROC                         | AUROC             |
|                               | **Reference, logistic regression** | **0.67 (0.66 — 0.68)** | **0.74 (0.71 — 0.75)** |
|                               | **Gradient boosting**          | **0.71 (0.70 — 0.73)** | **0.75 (0.73 — 0.77)** |
| Chronic kidney disease (incidence: 0.03) | AUROC                         | AUROC             |
|                               | **Reference, logistic regression** | **0.68 (0.66 — 0.70)** | **0.77 (0.75 — 0.78)** |
|                               | **Gradient boosting**          | **0.77 (0.75 — 0.78)** | **0.77 (0.75 — 0.79)** |

(a) one year after first T2D diagnosis

(b) two years after first T2D diagnosis

(c) three years after first T2D diagnosis

(d) four years after first T2D diagnosis

**Supplementary Table 8.** Gradient boosting and baseline model prediction performance for different dates of prediction. Incidence is the proportion of cases within the population. AUROC is area under receiver operating characteristic curve which confidence intervals were obtained by bootstrap sampling procedure. T2D, type 2 diabetes.
### Heart failure

| Model                        | AUROC<sub>ts</sub> | AUROC<sub>no-ts</sub> |
|------------------------------|--------------------|-----------------------|
| Reference, logistic regression | 0.74 (0.72 — 0.75) | 0.76 (0.74 — 0.77)    |
| Logistic regression          | 0.77 (0.76 — 0.79) | 0.78 (0.77 — 0.80)    |
| Random forest                | 0.77 (0.75 — 0.78) | 0.80 (0.79 — 0.81)    |
| Gradient boosting            | 0.80 (0.78 — 0.81) | 0.81 (0.80 — 0.83)    |

### Myocardial infarction

| Model                        | AUROC<sub>ts</sub> | AUROC<sub>no-ts</sub> |
|------------------------------|--------------------|-----------------------|
| Reference, logistic regression | 0.68 (0.65 — 0.70) | 0.69 (0.67 — 0.71)    |
| Logistic regression          | 0.70 (0.68 — 0.73) | 0.71 (0.69 — 0.73)    |
| Random forest                | 0.67 (0.64 — 0.69) | 0.71 (0.69 — 0.73)    |
| Gradient boosting            | 0.71 (0.69 — 0.73) | 0.72 (0.70 — 0.74)    |

### Stroke

| Model                        | AUROC<sub>ts</sub> | AUROC<sub>no-ts</sub> |
|------------------------------|--------------------|-----------------------|
| Reference, logistic regression | 0.71 (0.69 — 0.73) | 0.72 (0.70 — 0.73)    |
| Logistic regression          | 0.72 (0.70 — 0.74) | 0.72 (0.70 — 0.74)    |
| Random forest                | 0.69 (0.67 — 0.71) | 0.71 (0.69 — 0.73)    |
| Gradient boosting            | 0.72 (0.70 — 0.74) | 0.73 (0.72 — 0.75)    |

### Cardiovascular disease

| Model                        | AUROC<sub>ts</sub> | AUROC<sub>no-ts</sub> |
|------------------------------|--------------------|-----------------------|
| Reference, logistic regression | 0.66 (0.64 — 0.67) | 0.67 (0.66 — 0.68)    |
| Logistic regression          | 0.68 (0.67 — 0.69) | 0.69 (0.68 — 0.70)    |
| Random forest                | 0.68 (0.67 — 0.69) | 0.70 (0.69 — 0.71)    |
| Gradient boosting            | 0.69 (0.68 — 0.70) | 0.71 (0.70 — 0.72)    |

### Chronic kidney disease

| Model                        | AUROC<sub>ts</sub> | AUROC<sub>no-ts</sub> |
|------------------------------|--------------------|-----------------------|
| Reference, logistic regression | 0.71 (0.69 — 0.73) | 0.72 (0.69 — 0.73)    |
| Logistic regression          | 0.74 (0.72 — 0.76) | 0.75 (0.73 — 0.77)    |
| Random forest                | 0.74 (0.72 — 0.76) | 0.76 (0.74 — 0.78)    |
| Gradient boosting            | 0.77 (0.76 — 0.79) | 0.78 (0.76 — 0.80)    |

### All-cause mortality

| Model                        | AUROC<sub>ts</sub> | AUROC<sub>no-ts</sub> |
|------------------------------|--------------------|-----------------------|
| Reference, logistic regression | 0.80 (0.79 — 0.81) | 0.80 (0.79 — 0.81)    |
| Logistic regression          | 0.83 (0.82 — 0.83) | 0.83 (0.82 — 0.83)    |
| Random forest                | 0.85 (0.85 — 0.86) | 0.84 (0.84 — 0.85)    |
| Gradient boosting            | 0.87 (0.86 — 0.87) | 0.86 (0.86 — 0.87)    |

**Supplementary Table 9.** Comparison of model performance between time-split and non-time-split models. AUROC measure for each prediction model types best parametrization (according to AUROC measure) for all outcomes compared between time-split (AUROC<sub>ts</sub>; training, test and validation sets were split so that the model is trained on individuals diagnosed with type 2 diabetes historically earlier and evaluated on individuals diagnosed later) and non-time-split model (AUROC<sub>no-ts</sub>; training, test and validation sets were split at random without accounting for date of prediction). I<sub>ts</sub> and I<sub>no-ts</sub>, outcome incidence within time-split and non-time-split study populations, respectively. AUROC, area under receiver operating characteristic curve which confidence intervals were obtained through bootstrap sampling procedure.
Supplementary Table 10. Area under receiver operating characteristic curve for prediction of all-cause mortality by the best reference and register-based models. Compared to the prediction of T2D comorbidities, all models achieved relatively high AUROCs, with all register-based models outperforming the reference model and gradient boosting outperforming the other register-based models. T2D, type 2 diabetes; AUROC, area under receiver operating characteristic curve.

| Model                      | AUROC   | ΔAUROC<sub>RLR</sub> | ΔAUROC<sub>LR</sub> | ΔAUROC<sub>RF</sub> |
|----------------------------|---------|----------------------|--------------------|--------------------|
| Reference, logistic regression (RLR) | 0.80 (0.79 — 0.81) |                      |                    |                    |
| Logistic regression (LR)    | 0.83 (0.82 — 0.83) | 0.03 (0.02 — 0.03)   |                    |                    |
| Random forest (RF)          | 0.85 (0.85 — 0.86) | 0.05 (0.05 — 0.06)   | 0.03 (0.02 — 0.03) |                    |
| Gradient boosting (GB)      | 0.87 (0.86 — 0.87) | 0.07 (0.06 — 0.07)   | 0.04 (0.04 — 0.05) | 0.01 (0.01 — 0.02) |

Supplementary Figure 8. Death is a competing risk for diagnosing T2D comorbidities and its effect may be estimated by ML models implicitly. We investigated whether health register data and machine learning model could predict five-year risk of all-cause mortality (ACM) using the same procedure as for T2D comorbidities. (a) Five-year incidence of ACM for population percentiles ranked by ACM risk predicted by gradient boosting (blue) and reference (orange) models. At the 95th percentile, both models identified individuals ranking multiples above the overall population incidence (ACM risk ratio of 4.53, and 3.46, respectively). (b) Incidence of ACM among population percentiles according to predicted risk of chronic kidney disease. Individuals stratified by the register-based model had a similar or lower risk of ACM than those binned into top percentiles by the reference model. (c) 50 register features most predictive of ACM according to the gradient boosting models feature importances. Unlike the case of the investigated T2D comorbidities, hospital diagnoses features were the most important followed by prescriptions second and canonical features third.
Supplementary Figure 9. All-cause mortality five-year incidence in population percentiles ranked by predicted type 2 diabetes comorbidity risk by the best gradient boosting (green) and best baseline (violet) models.