SCIENTIFIC REPORTS
nature research

OPEN

Nationwide prediction of type 2 diabetes comorbidities

Piotr Dworzynski[1,2], Martin Aasbrenn[1,3], Klaus Rostgaard[2], Mads Melbye[2,4,5], Thomas Alexander Gerds[6], Henrik Hjalgrim[2,7,8] & Tune H. Pers[1,2,8*]

Identification of individuals at risk of developing disease comorbidities represents an important task in tackling the growing personal and societal burdens associated with chronic diseases. We employed machine learning techniques to investigate to what extent data from longitudinal, nationwide Danish health registers can be used to predict individuals at high risk of developing type 2 diabetes (T2D) comorbidities. Leveraging logistic regression-, random forest- and gradient boosting models and register data spanning hospitalizations, drug prescriptions and contacts with primary care contractors from >200,000 individuals newly diagnosed with T2D, we predicted five-year risk of heart failure (HF), myocardial infarction (MI), stroke (ST), cardiovascular disease (CVD) and chronic kidney disease (CKD). For HF, MI, CVD, and CKD, register-based models outperformed a reference model leveraging canonical individual characteristics by achieving area under the receiver operating characteristic curve improvements of 0.06, 0.03, 0.04, and 0.07, respectively. The top 1,000 patients predicted to be at highest risk exhibited observed incidence ratios exceeding 4.99, 3.52, 1.97 and 4.71 respectively. In summary, prediction of T2D comorbidities utilizing Danish registers led to consistent albeit modest performance improvements over reference models, suggesting that register data could be leveraged to systematically identify individuals at risk of developing disease comorbidities.

Comorbidities of type 2 diabetes (T2D) represent a major cause of death and disabilities resulting in substantial societal and economic burdens[1,2]. Early interventions have been shown to delay the onset of comorbidities in chronic disease, for instance, intensified multifactorial intervention has been successful in lowering risk of cardiovascular events and slowing down progression of renal disease in patients diagnosed with T2D and microalbuminuria[3]. However, due to the multifactorial underpinnings of chronic diseases and their high prevalence, population-wide interventions remain challenging and hampered by cost-effectiveness and safety concerns[4,5]. Recent studies[4–7] and health policy recommendations[2,8] suggest that these challenges – especially in the long run[9] – could be addressed by tailoring interventions to high-risk individuals. Thus, identifying sub-populations at the highest risk of developing chronic disease comorbidities could pave the way for improvements in resource allocation and health outcomes.

Machine learning (ML) techniques are increasingly being used to analyze electronic health record data to predict future disease onset or its future course[10–14]. These efforts include prediction of onset and complications of cardiovascular disease[15–22], onset of T2D[23–28], onset of kidney disease[29], as well as prediction of postoperative outcomes[30–34], birth related outcomes[35,36], mortality[16,37,38] and hospital re-admissions[39–45]. However, current approaches typically suffer from a number of limitations. First, although altering the course of chronic disease requires timely prediction[7], relatively few electronic health record-based studies employ long prediction horizons (the time window during which the studies predicted whether a given individual would be diagnosed with a particular outcome). For example, a recent review[10] noted that out of 81 relevant studies, only 32 used at least a one-year prediction horizon and only 15 used at least a five-year prediction horizon. Second, the above-referenced studies typically do not explicitly test whether the trained prediction model can predict future events[7]. As both medical and data aggregation practices change over time, models trained on data covering a relatively long time period are susceptible to learning patterns, which become irrelevant[46,47] - meaning that a model can perform

[1]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [2]Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark. [3]Department of Geriatrics and Internal Medicine, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark. [4]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. [5]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [6]Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark. [7]Department of Haematology, Rigshospitalet, Copenhagen, Denmark. [8]These authors contributed equally: Henrik Hjalgrim and Tune H. Pers. *email: tune.pers@sund.ku.dk

Register	Data	Int. classification	Dates	Total #records
DN Patient Register	Hospital diagnoses	ICD-10	1995 → 2016*	179 million
	Surgical procedures	NCSP	1996 → 2016	26.9 million
	Treatment procedures	—	1999 → 2016	103.4 million
	Diagnostic procedures	—	1999 → 2016	62.5 million
DN Prescription Register	Drug prescriptions	ATC	1994 → 2016	930.4 million
DN Health Service Register	Claims data	—	1990 → 2016	2,147 million
DN Medical Birth Register	Pregnancies & prenatal care	—	1973 → 2016	2.7 million

Table 1. Overview of Danish national health registers employed in this study. *ICD-8 used between 1977 and 1995, ICD-10 between 1995 and 2016. ICD-10 stands for 10th revision of International Statistical Classification of Diseases and Related Health Problems; DN, Danish National; NCSP, Nordic Medico-Statistical Committee Classification of Surgical Procedures; ATC, Anatomical Therapeutic Chemical Classification; #, number of.

poorly on new data despite having a good average performance on the source data. Lastly, the use of selected cohorts that are not representative of the general population can have important limitations. By relying on data from a specific place of care, state or electronic health record system, the ML model can develop biases specific to that data[7,10] and generalize poorly when used for other populations. Overcoming these limitations requires not only an appropriate methodological approach but, importantly, a dataset that is informative, to allow for accurate prediction; long-term, to enable a sufficient prediction horizon and modelling of temporal trends in data; and generalizable, to ensure that the population on which the model is trained is sufficiently representative of the population on which the model is to be applied on.

Most countries do not maintain centralized, nationwide health registers suitable for electronic health record analyses. A notable exception is Denmark, which stores national longitudinal health register data including population-wide information on drug prescriptions, hospital diagnoses, hospital medical procedures as well as claims information from primary care contractors of which all, besides the latter, use international code classifications[48,49]. For example, there are 348,000 individuals diagnosed with T2D from 1995 to 2011. For these individuals, the registers contain information on >14.2 million hospital diagnoses (>15,900 unique codes), >21.3 million procedures (>13,000 unique), >182 million claims from primary care contractors (>12,600 unique) and >143.7 millions prescriptions (>22,000 unique) amounting to a total of >553.2 million medical events (defined as time-stamped register entries potentially predictive of some future outcome event). In sum, the Danish longitudinal registers contain the approximate medical life course of virtually every Danish citizen, possibly presenting valuable opportunities for predicting adverse medical outcomes, such as comorbidities in the chronically ill.

Here, we leveraged Danish register data and commonly used ML methods to assess the extent to which nationwide longitudinal records on hospital diagnoses, hospital procedures, prescriptions and claims from primary care contractors covering 10 years of nearly >200,000 T2D patients, could predict future onset of chronic disease comorbidities in these patients. We chose to focus on T2D due to its prevalent nature, its diverse commodities and because early intervention has been shown to reduce subsequent risk of developing commodities. Applying three common ML methods, namely logistic ridge regression, random forest and gradient boosting we predicted future onset of MI, HF, CVD, CKD and stroke within a five-year window of individuals' first T2D diagnosis.

Materials and Methods

Danish population-wide registers. The Danish National Patient Register[50] contains information on all non-psychiatric hospital visits since 1977 and includes psychiatric hospital, emergency department and outpatient clinic visits since 1995. All medical events are recorded using the Health Care Classification System, which combines multiple Danish and international classifications. All diagnoses are recorded using the International Statistical Classification of Diseases and Related Health Problems (ICD) system (revision 8 used between 1977 and 1994 and revision 10 used since 1995); surgeries from 1996 onwards are recorded using the Nordic Medico-Statistical Committee Classification of Surgical Procedures (NCSP); treatment and diagnostic procedures are represented through unique Danish hierarchical classifications. The Danish National Prescription Register[51] contains detailed information on all prescription drugs sold in Danish pharmacies since 1995 and uses the Anatomical Therapeutic Chemical (ATC) classification system. The Danish National Health Service Register contains claims information on "activities of health professionals such as general practitioners, practising medical specialists, physiotherapists, dentists, psychologists, chiropractors, and chiropodists"[52] and uses a unique classification scheme. Furthermore, we included data from the Central Person Register comprising an individual's sex, date of birth, country or region of birth, date of death as well as generalized residence address (five areas throughout Denmark). The pregnancy and childbirth information used to identify gestational diabetes was extracted from The Danish Medical Birth Register[53]. Causes of death were identified through the Danish Register of Causes of Death[54]. More detailed information regarding the registers can be found in Table 1.

Study population and inclusion criteria. We employed an observational prediction study with a five-year prediction horizon using as the time of prediction the date of an individual's first T2D diagnosis. We selected a population of (n = 203,517) T2D patients. The date of the first T2D diagnosis was defined as the date of first registration of a T2D hospital diagnosis (ICD-10 code E11), prescription of insulin and analogues (ATC code A10A) or prescription of blood glucose lowering drugs (ATC code A10B). As precise dates of hospital diagnoses

	T2D population	Heart failure	Myocardial infarction
# Individuals	203,517	190,819	190,987
# Cases	—	8,940 (4.69%)	6,485 (3.40%)
# Non-cases	—	181,879 (95.31%)	184,502 (96.60%)
% Women	47.03	47.52	48.14
Median age at T2D diagnosis	61.44	60.67	60.95
Median # days until outcome	—	1631.60	1537.80
# Features in RFV	—	6,155	6,155
	Stroke	**Cardiovascular disease**	**Chronic kidney disease**
# Individuals	191,095	120,114	200,646
# Cases	7,922 (4.15%)	33,057 (27.52%)	5,617 (2.80%)
# Non-cases	183,173 (95.85%)	87,057 (72.48%)	195,029 (97.20%)
% Women	47.38	49.77	47.18
Median age at T2D diagnosis	60.80	57.25	61.32
Median # days until outcome	1641.60	1414.00	2065.30
# Features in RFV	6,155	6,155	6,155

Table 2. Overview of study population characteristics among all newly diagnosed type 2 diabetics (T2D population) and five comorbidity populations (individuals undiagnosed with a comorbidity at the time of their first T2D diagnosis). T2D, type 2 diabetes; RFV, register feature vector; #, number of.

were not known, we chose to use the last day of the relevant hospital admission as the T2D diagnosis date and the first day of the relevant hospital admission as the outcome diagnosis date. We excluded A10B prescriptions that occurred during pregnancy as well as individuals who had their first A10A prescription before the age of 30. We excluded individuals for whom the date of prediction occurred prior to January 1 2000 and after December 31 2010 to prevent inflation of first-T2D-diagnoses during the first years of register information availability and to establish five-year follow-up for all patients. Thus the length of time windows to ensure that there indeed was no prior T2D diagnosis ranged from five to 16 years. Additionally, to evaluate the predictiveness of medical events following the first T2D diagnosis, we also selected four populations of T2D patients with time of prediction set to one year (n = 181,100), two years (n = 165,482), three years (n = 150,228), and four years (n = 137,240) after the first T2D diagnosis (maintaining the five-year prediction horizon).

For each of the five T2D comorbidities the prediction objective was the five-year risk of an individual's first diagnosis of the specific comorbidity. Specifically, we chose the first occurrence of a non-referral ICD-10 hospital diagnosis code within the Danish National Patient Register or ICD-10 code assignment as cause of death in the Danish Register of Causes of Death related to HF (ICD-10 code H50), MI (ICD10 code I21), ST (ICD-10 code I61-I64), CVD (ICD-10 codes I) and CKD (ICD-10 code N17-N19). Prior validation of top-level ICD-10 diagnosis codes in the Danish National Patient Register has shown overall high precision of the codes we used to define comorbidities, positive predictive values ranged from 98.0 to 100 for acute MI, 80.8 to 100 for HF, 79.3 to 93.5 for stroke[50]. For each comorbidity, a case was defined as an individual with a first-time diagnosis of the given comorbidity occurring within the five-year prediction horizon, and similarly, a non-case was defined as an individual with no diagnosis of the given comorbidity occurring during the prediction horizon. Consequently, individuals who died during the prediction horizon period without prior comorbidity diagnosis were treated as non-cases. Considerations on how all-cause mortality may impact the prediction models can be found in Supplementary Note 1. To avoid instances where the comorbidity was diagnosed before the diagnosed onset of T2D (for example through an informative diagnostic test ordered to confirm prior beliefs about a given outcome), we required comorbidities to be diagnosed at least 30 days after the diagnosed date of T2D onset (henceforth *buffer period*). This restriction effectively set the time zero of the prediction to 30 days after the individuals' first T2D diagnosis. The length of the buffer period was chosen based on the length distribution of hospital admissions during which individuals with a hospital-based T2D diagnosis received their first T2D diagnosis (Supplementary Fig. 1). We chose 30 days because that threshold captured the majority of the relatively short hospital admissions (73.5% of the hospital-based T2D diagnoses stemmed from this time interval). Confirmatory prediction analysis for the CKD comorbidity with the buffer period length set to 60 days led to no notable differences in population characteristics nor model performances (Supplementary Tables 1 and 2). For a study overview and study population characteristics, please refer to Table 2, Fig. 1 and Supplementary Table 3.

Predictor variables. For all individuals, we created two sets of predictive variables (henceforth *feature vectors*): a canonical feature vector, composed of canonical features, to be used by a reference model, and a register feature vector, composed of canonical features as well as features extracted from the Danish health registers, to be used by the ML models. The canonical features, chosen based on their readily population-wide availability and known relationship to comorbidity risk, were sex (binary encoded), country or region of birth (as denoted in the Central Person Register; 13 categories), date of prediction (continuous), age at date of prediction (continuous), number of days lived in a general region of Denmark (continuous; five regions), an interaction term between age and sex, a restricted cubic spline of the prediction date and a restricted cubic spline of age at date of prediction. The register features denoted counts of individuals' hospital diagnoses, drug prescriptions, procedures and interactions with primary care contractors. Only non-referral diagnoses encoded in the ICD-10 system were included.

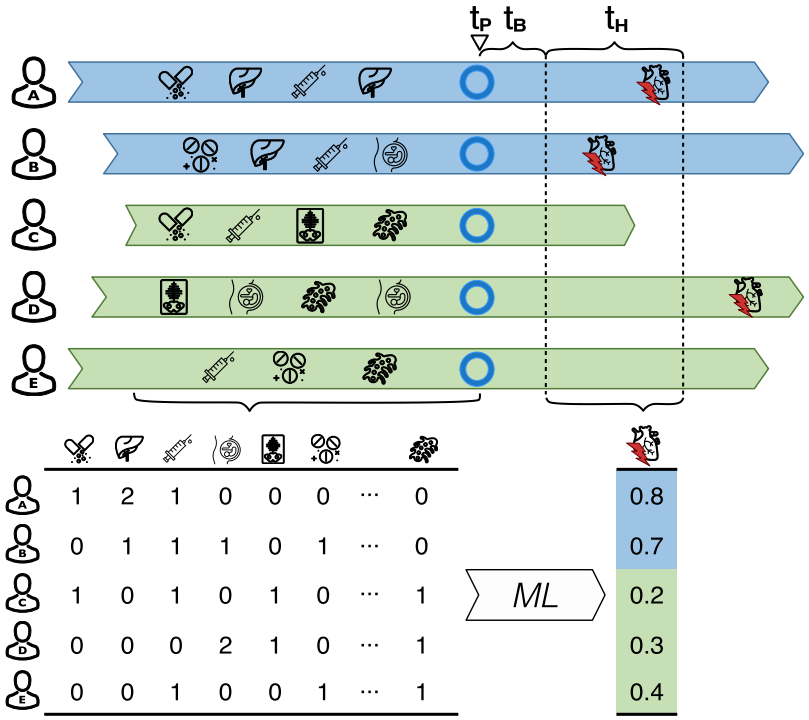

Figure 1. The colored arrows represent each individual's accumulated history of register events (determinants, used as predictive features). t_p represents the date of prediction (in this case the day of the first T2D diagnosis represented by blue circles). t_B depicts the buffer period of 30 days used to exclude individuals who were diagnosed with the comorbidity shortly before t_P. t_H represents the five-year prediction horizon. Individuals for whom the comorbidity occurred before the prediction horizon were removed. Individuals with their first comorbidity diagnosis occurring within the prediction horizon are referred to as *cases* and all others are referred to as *non-cases*. For each individual, register features representing medical events which occurred before the time of prediction are aggregated into counts (table below) and used as the prediction model's predictor variables (features) to determine the likelihoods of being a case or a non-case. ML, machine learning.

Claims from primary care contractors were encoded as a count of a given individual's interactions with a specific health care contractor type (e.g. "general practitioner" or "dentist"). Each diagnosis, drug prescription, and medical procedure was included at multiple levels of specificity by leveraging their respective code hierarchies (ICD-10 for diagnoses, NCSP for procedures and ATC for drug prescriptions). For example, the diagnosis "other type of myocardial infarction" (ICD-10 code I21.A) was represented by three features, namely diagnosis of "diseases of the circulatory system" (ICD-10 code I), diagnosis of "acute myocardial infarction" (ICD-10 code I21) and diagnosis of "other type of myocardial infarction" (ICD-10 code I21.A). For a given individual the value of each feature denoted the cumulative number of observations of electronic health record codes representing that feature prior to the time of prediction. The register feature vector for all comorbidities comprised 6,181 register features: 26 canonical features, 3,423 hospital diagnoses, 2,015 hospital procedures, 670 prescriptions and 47 primary care interactions.

Prediction models. For each comorbidity (and all-cause mortality), four algorithms were used to predict an individual's risk of developing the comorbidity within the prediction horizon. Logistic ridge regression[55] with the canonical feature vector as input was used as the reference model. Restricted cubic splines were used to model non-linear effects of age separately for the two genders by adding an interaction term. As our register-based models, we used three algorithms (model types), namely logistic ridge regression, random forest[56], and decision tree gradient boosting[57] with the register feature vector as input. Briefly, logistic ridge regression is a logistic regression model where parameter estimates are shrunk towards zero by means of a penalty term (tuning: penalty parameter, maximum number of iterations); random forest is an ensemble model which in our case consists of many classification trees built in many bootstrap samples (tuning: number of trees, maximum depth of each tree); decision tree gradient boosting results in a sequence of shallow decision trees where each consecutive tree is trained to improve the performance of the ones before it (tuning: learning rate, maximum depth of each tree).

Dataset split and model training/evaluation procedures. For each comorbidity (and all-cause mortality), each prediction model was trained and evaluated using the following procedure. First, all individuals were split into three sets, namely a training set, a test set, and a validation set. The training set was allocated the first 70% of the individuals (sorted by date of first T2D diagnosis) of the full dataset and used for three-fold cross-validated[58] hyperparameter search. Following Corey *et al.*[30], the remaining (most recently diagnosed) 30%

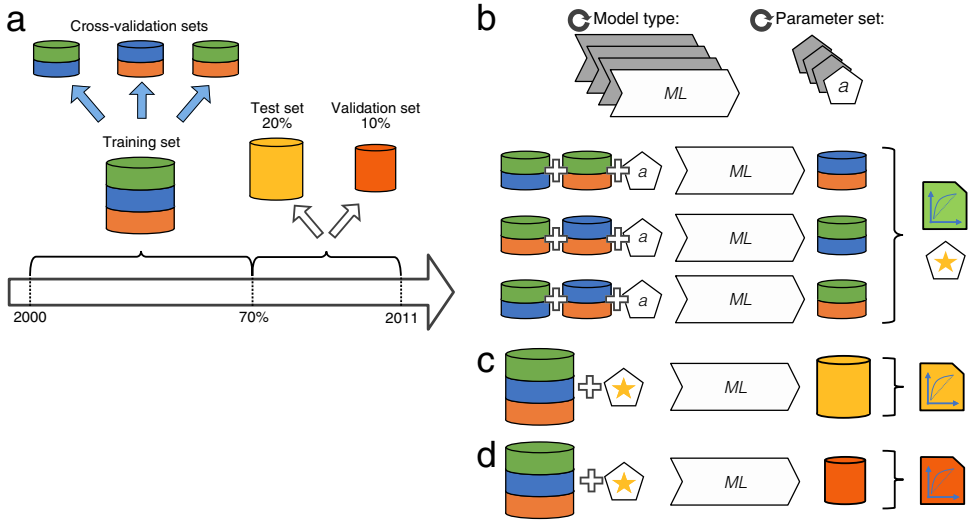

Figure 2. Data split and model tuning. (**a**) Data were split so that the training set constituted the first 70% of the data (time-wise, according to the time of the first T2D diagnosis). The test and validation sets were divided by balanced random sampling from the remainder (time-wise latter part) of the dataset. (**b**) For each model type, the best parameter set was chosen through its evaluation following three-fold cross-validation on the training data. (**c**) For each model type, the best model was obtained by re-training the model on the entirety of the training set using the best parameter set. Performance of each best model was evaluated on the test set for the purpose of development of this work. (**d**) Performance of each best model was evaluated on the validation set for the purpose of reporting the results. ML, machine learning.

of the full dataset were randomly split into two balanced (the proportion of cases was maintained in each split) subsets: a test set which was allocated 20% of the full dataset to be used for model selection for each outcome and a validation set which was allocated the remaining 10% of the full dataset and exclusively used to present final results. To improve training performance on the class unbalanced data (proportion of cases to controls ranging from 2.8% for CKD to 27.5% for CVD) models were trained with an algorithm-specifc class weighting regime (where training importance of each sample scaled by the inverse of its class prevalence). For each model type, the best parameter set was chosen by first repeatedly training the model on two-thirds of the training data and evaluating it on the remaining third, and then averaging accuracy measures across the three runs. The parameterization resulting in the highest average area under the receiver operating characteristic curve (AUROC) was used to re-train the model on the combined training set yielding the best model for a given model type. Each best model was then calibrated using the Platt Scaling method[59,60] on the test set. The best model's performance was evaluated against the test set and compared to the corresponding best models from the other algorithms. Finally, the validation set was used to report the calibrated best model's performance (reporting AUROC) and calibration error[61]. 95% confidence intervals of AUROC and AUROC differences (ΔAUROC) between models were obtained using the bootstrap method[62] using 1000 samples. The training and validation procedure is illustrated in Fig. 2 and an overview of all tested parameters can be found in Supplementary Table 4. The analysis was implemented in Python programming language[63] version 3.6.2 using modules numpy[64] version 1.13.1, pandas[65] version 0.20.3, scikit-learn[66] version 0.19.0, patsy[67] version 0.4.1, xgboost[68] version 0.80 and plots were made using bokeh visualization library[69] version 1.0.4. The source code is available at https://github.com/perslab/dworzynski-2020.

Ethical approval. The project was approved by Danish Data Protection Agency (2015-57-0102). Under Danish law register-based research does not require informed consent (section 10 of "The Act on Processing of Personal Data (DK)", Danish law nr 502 as of 23/05/2018).

Results

Prediction of T2D comorbidities based on medical events incurred prior to the first T2D diagnosis.

To assess to what extent register data can predict comorbidities for T2D patients, we first identified all individuals either diagnosed with T2D or being prescribed diabetes drugs between 2000 and 2011 (See Material and Methods; Fig. 1). Among these 203,517 individuals, we next identified subsets having received a hospital diagnosis of a major T2D comorbidity within a five-year period following their first T2D diagnosis. We identified 8,940 incident cases for HF, 6,485 for MI, 7,922 for ST, 33,057 for CVD and 5,617 for CKD, and treated the remaining individuals as non-cases (see Table 2 for study population characteristics).

The median number of medical events used for the prediction per individual was 381 for HF, 389 for MI, 391 for ST, 268 for CVD, 408 for CKD. For example, for CVD, each individual had an average of 33 hospital diagnosis events (the most common being essential hypertension, ICD-10 code I109), 78.5 hospital procedures (the most common being x-ray examination, SKS code UXR), 6.68 claims data events (the most common being contact

with general practitioner, SSR category 80), 420.5 drug prescriptions (the most common being antibacterials for systemic use, ATC code J01) prior to their first T2D diagnosis.

Using the AUROC measure, we first assessed how well we could predict the five T2D comorbidities based on medical events incurred prior to the individuals' first T2D diagnosis. The gradient boosting model yielded the highest AUROC for all outcomes and outperformed the reference model for four comorbidities (HF, MI, CVD, CKD; Table 3). The best gradient boosting models performed better than the best random forest models for all comorbidities and logistic ridge regression for HF, CVD and CKD. Additionally, register-based logistic ridge regression outperformed the reference model for HF, MI, CVD, and CKD. Together these results indicate that prediction based on hundreds of features assembled across health registers improves on prediction based on a limited number of canonical measures typically used to assess risk. Moreover, our results suggest that gradient boosting improves on logistic ridge regression models' performance in the context of large-scale health register data-based prediction. For the best models across all comorbidities the maximum expected calibration errors were relatively high (ranging from 0.18 for CVD to 0.45 for CKD). Calibration curves and error plots of the best models prior to and following calibration can be found in Supplementary Figs. 2–5. The following analyses were focused on predicting the HF, MI, CVD and CKD comorbidities (see Supplementary Fig. 6 for results on ST). Detailed results for each tested model parametrization can be found in Supplementary Tables 5 and 6.

Comorbidity incidences across predicted risk percentiles. To investigate the models' ability to rank individuals according to their risk of being diagnosed with a given comorbidity, we - for each comorbidity - grouped individuals into risk percentiles (based on the predicted risk from the gradient boosting and reference models) and plotted for each percentile the five-year comorbidity incidence (i.e. the fraction of individuals at that percentile incurring a given comorbidity during the prediction horizon; Fig. 3). Individuals in the highest percentiles for both the reference and register-based models exhibited markedly increased comorbidity incidences compared to the rest of the population. For example, for the gradient boosting model, individuals predicted to be in the 95th percentile had incidence risk ratios of 3.67 for HF, 2.66 for MI, 1.46 for CVD and 4.18 for CKD. Individuals predicted to be in the top gradient boosting-based risk percentiles had modest, but consistently higher five-year comorbidity incidences compared to individuals grouped by the reference model into the corresponding percentiles. This finding underscores that prediction based on multiple register-derived events is likely to improve on models based on a more limited number of canonical features.

A potential application of population-wide risk prediction models is to identify individuals at risk of a certain comorbidity (or another outcome). To further investigate whether individuals predicted to be at high risk in fact tended to be diagnosed with the given comorbidity, we computed for different thresholds at the right end tail of the predicted risk distributions, five-year comorbidity incidences for the best gradient boosting and reference models. At each threshold, we calculated a risk ratio, by comparing the five-year incidence of each of the four comorbidities for all individuals above that threshold (e.g. the proportion of cases to non-cases among the 1000 most at-risk predicted individuals) to the T2D patient population-wide incidence (Fig. 4). For the gradient boosting model of HF, CVD and CKD, we observed that the risk ratios were significantly higher than those based on the reference models (confidence intervals widened towards the rightmost end of the distribution due to the decreasing number of individuals for smaller thresholds).

Predictive features. To gain insight into the decision-making process of the best models' predictions, we analyzed the gradient boosting models' feature importance, i.e. the underlying features' relative contribution to the prediction. First, we assessed the contribution of each feature type (canonical features, prescriptions, address information, hospital diagnoses, and procedures, primary care interactions) for the 50 most predictive features as well as their cumulative importance (sum of all feature importance for each feature type; Fig. 5). We found that feature importance formed a long right-skewed distribution, indicating that model predictions were based on a large number of features. Furthermore, we identified prescription features as being the most important for all comorbidities (feature importance of 30% for ST to 41% for HF), followed by either canonical features (17% for HF to 27% for ST) or hospital diagnoses (20% for CKD to 27% for CVD). Detailed information on the first seven most important features for each outcome as well as their distribution in cases and non-cases can be found in Supplementary Fig. 7. Together, these results show that data from health registers, especially prescriptions and hospital diagnoses, provide predictive information on top of canonical features such as age, sex, and date of diagnosis, which are known to be predictive for the given comorbidities.

Inclusion of medical events incurred after the first T2D diagnosis does not improve prediction. We next investigated whether the inclusion, as predictive features, of medical events following the first T2D diagnosis could further improve prediction accuracy. First, we employed the same prediction procedure with the time of prediction set to one-, two-, three- and four years after the first T2D diagnosis while keeping the length of the buffer period (30 days) and prediction horizon (five years) constant. Contrarily to our expectation, we observed largely unchanged model performances in terms of AUROC (Supplementary Table 7). These results indicate that medical events succeeding the first T2D diagnosis did not contain new information predictive of the T2D comorbidities, suggesting that register-based prediction models are equally accurate when predicting T2D comorbidities at the date of first T2D diagnosis and during the following four years. For information on population characteristics and model performances please refer to the Supplementary Tables 3 and 8.

Comparison between time-split and non-time-split prediction. A commonly criticized aspect of register-based prediction analysis is that models are evaluated on data covering the same time period as the training data[7,8]. In such frameworks it is difficult to determine whether data patterns learned by the models are

	Heart failure (incidence: 0.04)			
	AUROC	$\Delta AUROC_{RLR}$	$\Delta AUROC_{LR}$	$\Delta AUROC_{RF}$
Reference, logistic regression (RLR)	0.74 (0.72–0.75)			
Logistic regression (LR)	0.77 (0.76–0.79)	0.04 (0.02–0.05)		
Random forest (RF)	0.77 (0.75–0.78)	0.03 (−0.01)	−0.01 (−0.02–0.01)	
Gradient boosting (GB)	0.80 (0.78–0.81)	0.06 (0.05–0.07)	0.02 (0.01–0.03)	0.03 (0.02–0.04)
	Myocardial infarction (incidence: 0.02)			
	AUROC	$\Delta AUROC_{RLR}$	$\Delta AUROC_{RL}$	$\Delta AUROC_{RF}$
Reference, logistic regression (RLR)	0.68 (0.65–0.70)			
Logistic regression (LR)	0.70 (0.68–0.73)	0.03 (0.01–0.04)		
Random forest (RF)	0.67 (0.64–0.69)	−0.01 (−0.03–0.01)	−0.04 (−0.06–−0.02)	
Gradient boosting (GB)	0.71 (0.69–0.73)	0.03 (0.02–0.05)	0.01 (0.00–0.02)	0.04 (0.03–0.06)
	Stroke (incidence: 0.03)			
	AUROC	$\Delta AUROC_{RLR}$	$\Delta AUROC_{RL}$	$\Delta AUROC_{RF}$
Reference, logistic regression (RLR)	0.71 (0.69–0.73)			
Logistic regression (LR)	0.72 (0.70–0.74)	0.01 (0.00–0.01)		
Random forest (RF)	0.69 (0.67–0.71)	−0.02 (−0.04–−0.01)	−0.03 (−0.04–−0.01)	
Gradient boosting (GB)	0.72 (0.70–0.74)	0.01 (0.00–0.02)	0.01 (0.00–0.02)	0.03 (0.02–0.05)
	Cardiovascular disease (incidence: 0.25)			
	AUROC	$\Delta AUROC_{RLR}$	$\Delta AUROC_{LR}$	$\Delta AUROC_{RF}$
Reference, logistic regression (RLR)	0.66 (0.64–0.67)			
Logistic regression (LR)	0.68 (0.67–0.69)	0.02 (0.02–0.03)		
Random forest (RF)	0.68 (0.67–0.69)	0.02 (0.02–0.03)	0.00 (0.00–0.01)	
Gradient boosting (GB)	0.69 (0.68–0.70)	0.04 (0.03–0.05)	0.02 (0.01–0.02)	0.01 (0.01–0.02)
	Chronic kidney disease (incidence: 0.03)			
	AUROC	$\Delta AUROC_{RLR}$	$\Delta AUROC_{LR}$	$\Delta AUROC_{RF}$
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)

Table 3. AUROC measures for each prediction model's best parameterization. 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 five T2D comorbidities. 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. For heart failure, myocardial infarction, cardiovascular disease and chronic kidney disease the gradient boosting model outperformed the reference models. AUROC, area under receiver operating characteristic curve.

predictive for future events due to potential changes in medical practice or changes in health patterns in general. In this work, we applied a so-called *time-split* approach (training set comprising of the first 70% of the population based on date of first T2D diagnosis, with the test and validation sets randomly sampled from the remaining 30%). To test whether sampling of the training, test and validation sets without taking into account individuals' dates of first T2D diagnoses, would yield different results, we re-ran all analysis with a *non*-time-split approach. Expectedly, the AUROCs from the non-time-split approach were consistently, although not significantly, better than AUROCs from the time-split approach applied here (Supplementary Table 9). Thus, we could not show that medical practice predictive of comorbidities has changed between January 1 2000 and December 31 2010, however, the consistent differences underline the potential bias of models trained in a non-time-split approach towards overestimating prediction accuracies of future events.

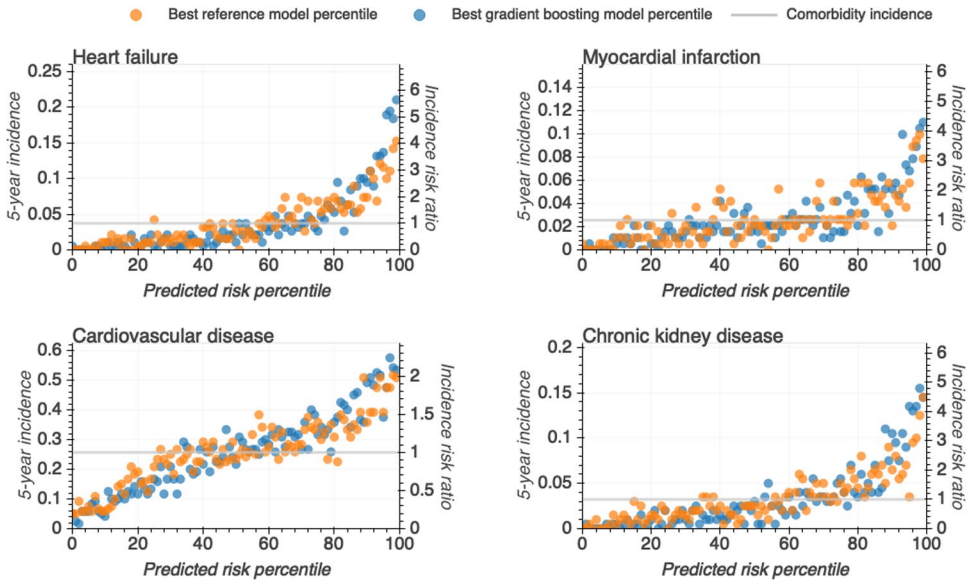

Figure 3. All individuals were ranked according to their predicted risk (in increasing order) by the best gradient boosting (blue) and best reference (orange) models and binned into percentiles. Plotted are the observed five-year comorbidity incidences for individuals in each percentile. Left y-axis; incidence defined as the observed proportion of individuals who did develop the given comorbidity within the five-year prediction horizon. Right y-axis; the incidence risk ratio defined as a ratio between the percentiles' and population observed five-year comorbidity incidence. Gray horizontal line; population five-year comorbidity incidence.

Discussion

Here we evaluated the extent to which commonly used ML methods when applied on comprehensive Danish nationwide health register data could predict selected T2D comorbidities, namely heart failure (HF), myocardial infarction (MI), stroke (ST), cardiovascular disease (CVD) and chronic kidney disease (CKD). Using the AUROC performance measure, we compared register-based models - logistic ridge regression, random forest and gradient boosting applied to nationwide information on hospital diagnoses, hospital procedures, drug prescriptions and claims information from primary care contractors - to a reference logistic ridge regression model based on age, sex, date of first T2D diagnosis and their interactions. For all comorbidities except ST, the register-based models outperformed the reference model. The gradient boosting model yielded the highest AUROCs significantly outperforming register-based logistic ridge regression for HF, CVD, and CKD as well as outperforming random forest models for all outcomes. For all register-based models, feature importances formed a long-tailed right-skewed distribution with the most important feature types being prescriptions, followed by the canonical features and hospital diagnoses. These results suggest that the predictive advantage of the register-based gradient boosting model stems from both the wide inclusion of predictive features from health registers, especially drug prescriptions, as well as increased flexibility (capacity to learn complex, non-linear patterns in data) of the gradient boosting model over logistic ridge regression models.

The predictive advantage of the register-based gradient boosting models over the register-based logistic ridge regression stands in contrast to recent findings suggesting an absence of significant improvements from using ML techniques over logistic regression in clinical prediction tasks[14]. We believe that this improvement, as well as applicability of ML approaches on electronic health record data generally, depend on at least three factors. First, on the number of predictive features - our study used 6,181 features while the median number of features for studies presented by Christodoulou *et al*. was 19. Second, the presence of non-linear predictive patterns within the data may favor random forest and gradient boosting over logistic regression models. For example, in our study, the register-based gradient boosting did not outperform logistic ridge regression only for those outcomes (MI and ST) for which general trends, as explained by canonical predictive features (age, sex, date of diagnosis), were the most predictive (cumulative importance of 26% and 27%, respectively). Lastly, the number of samples sufficient for an ML model to learn the aforementioned complex patterns (the number of individuals in our study ranges from 120,114 to 200,646 as compared to 1,250 samples for the studies presented by Christodoulou *et al.*). Overall, these findings demonstrate that employment of ML techniques should only follow an earnest effort in application of simpler modelling techniques, such as logistic regression, as the potential performance gains from ML approaches may not outweigh the associated loss of interpretability or increased risk of over-fitting.

We note that the predictive performance of all models was relatively modest. The best AUROC performances were observed for gradient boosting for HF achieving an AUROC of 0.80 and CKD with an AUROC of 0.77, in both cases outperforming the reference model by $\Delta\text{AUROC} = 0.05$. Overall, the reference model achieved AUROCs from 0.62 for CVD to 0.75 for HF (95% CI) while gradient boosting achieved AUROCs from 0.69 for MI up to 0.81 for HF (95% CI). Focusing on the individuals predicted to be at highest risk, the register-based models were able to identify patients whose five-year T2D comorbidity risk was considerably increased, with

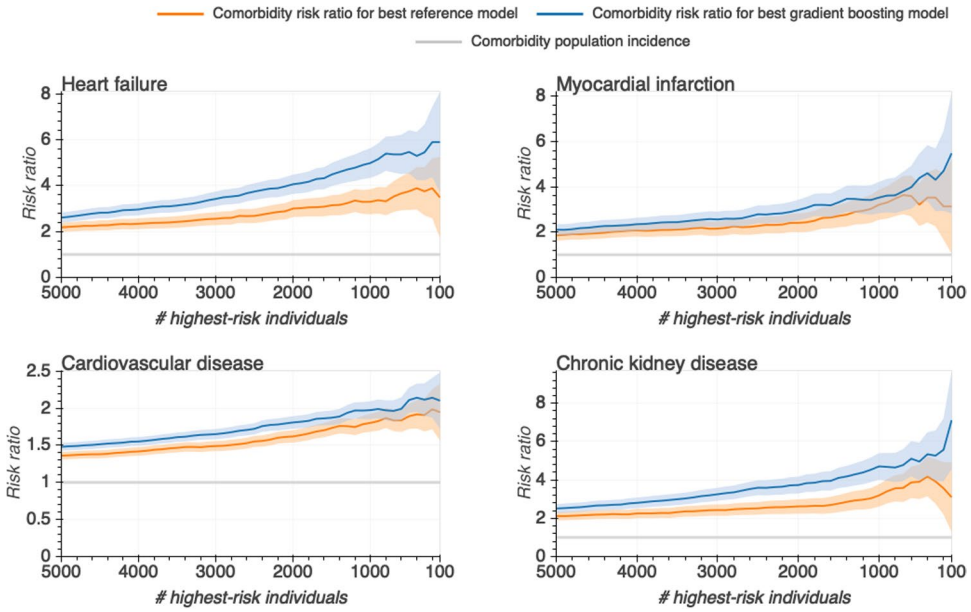

Figure 4. For each comorbidity individuals were ranked according to their predicted risk by the gradient boosting (blue) and reference (orange) models. For a number of individuals predicted to have the highest risk, risk ratios were calculated as the comorbidity incidence of individuals ranking above that threshold over the comorbidity incidence in the entire study population. 95% confidence intervals (shaded areas) were obtained through bootstrap sampling.

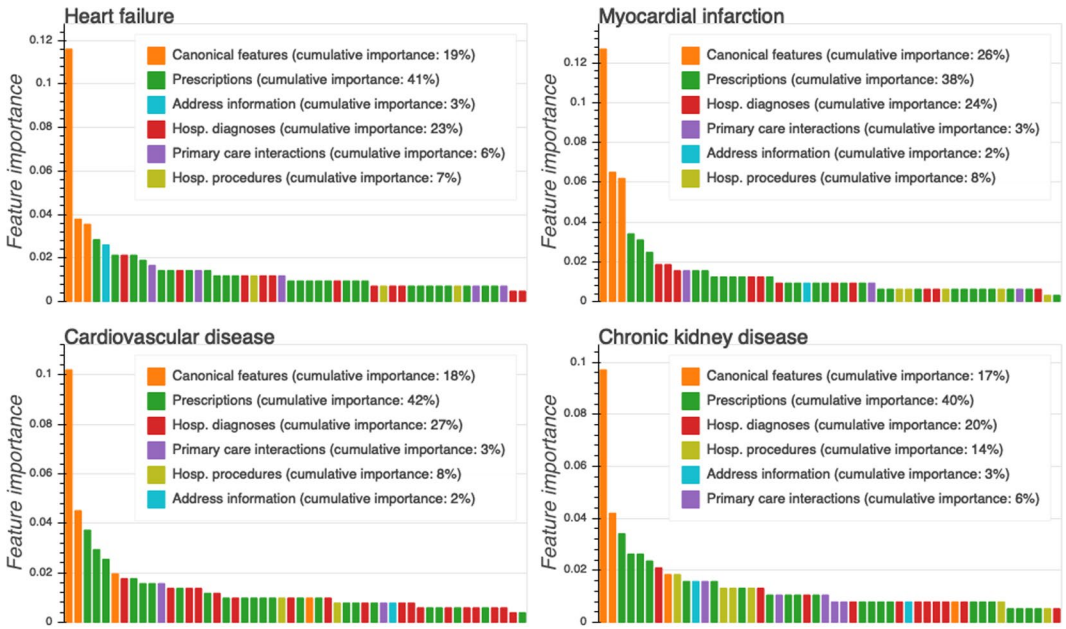

Figure 5. Top 50 most predictive register feature vector features from the best gradient boosting model ranked by their importance and colour-coded according to type (x-axis). Feature importance is a normalized estimate of a relative contribution of the feature to the model prediction (y-axis). Drug prescription features had the highest overall cumulative importance followed by the canonical features and hospital diagnoses. Age, interaction between age and sex, and date of first T2D diagnoses were the three most important features for all comorbidities.

risk ratios exceeding 3.1 for CVD, 3.8 for ST, 4.2 for MI, 4.8 for CKD and 5.5 for HF, which all were higher than the corresponding values for the reference model. Overall these results suggest that ML methods, such as gradient boosting, may outperform logistic regression in prediction on data from nationwide health registers. Furthermore, with continued development and careful consideration, such an approach could be used to

prioritize individuals for treatment regimens or to identify sub-groups of patients for whom an intervention could be cost-effective.

We found that AUROC values for prediction of comorbidities one-, two-, three- and four years after the first T2D diagnosis were comparable to prediction at the date of the first T2D diagnosis. This suggests that predictive register-derived risk factors for the investigated T2D comorbidities may already be present prior to a given individual's first T2D diagnosis. Similarly, while we observed a consistent increase in AUROCs without using a time-split approach, these increases did not reach statistical significance, suggesting that changes in medical practice between 2000 and 2016 had only modest impact on the comorbidity diagnoses during that period. However, in our opinion, using a time-split approach is a more correct and principled way of evaluating prediction models in healthcare settings. For example, recent changes in medical practice, such as increased use of statins[70] or introduction of new types of insulin or glucose-lowering drugs[71] in the period of 2000 to 2013 could reduce the relevance of patterns learned from earlier predictions[46] and, without proper evaluation, lead to overestimation of the models' abilities to predict future events.

Employment of Danish nationwide health register data for prediction of health-related outcome has a number of advantages. Firstly, the registers cover virtually every Danish citizen, which reduces the risk of training models based on unequal representations of population sub-groups and minimizes censoring. Secondly, our analyses include information from multiple Danish registers spanning from hospitalization events to drug prescriptions and claims from primary care contractors. Thus, our results give a more realistic picture of the extent to which register data can aid in the prediction of health-related outcomes, in our case T2D comorbidities. The observation that the register-based models exhibited consistently higher performances compared to the reference models underlines that register-based models may represent powerful frameworks to identify individuals at high risk of developing T2D comorbidities. Taking into account eligibility criteria, our approach can be used to prioritize individuals for treatment regimens or to identify sub-groups of patients for whom an intervention could become more cost-effective. Furthermore, due to the nationwide breadth of this data, our models could be retrained with genetics data, socioeconomic information, lab measurements and/or other comprehensive molecular data types and electronic medical record systems to move a step closer towards being incorporated into the routine care of patients newly diagnosed with T2D. The proof-of-concept approach presented in this work could be well-suited for a prediction of outcomes secondary to a wide array of diseases including cardiovascular and hematological diseases and potentially cognitive and psychiatric disorders.

Despite its strengths, our work has a number of important shortcomings. Firstly, due to the lack of diagnoses information from general practitioners, we relied on hospital diagnoses to identify the onset of investigated T2D comorbidities. Thus, we possibly included as cases individuals who already had been diagnosed by their general practitioner and potentially already were receiving treatment at the time of prediction. Our models could then identify the treatment-related prescriptions as highly predictive of a future hospital diagnosis of the comorbidity. This may be especially true for CVD, for which we observed a prescription of cardiovascular drugs as one of the most predictive features. Similarly, some of the individuals predicted to be at highest risk are already known to be in the high-risk group. For example, for CKD the best model in part relied on prescriptions of calcium blockers - agents typically prescribed to individuals with hypertension who because of this diagnosis are at risk of kidney disease. This indicates that some of the at-risk individuals were known to be at risk for CKD before their first T2D diagnosis. In future work, these confounders could be addressed by filtering out individuals with medical prescriptions indicative of the comorbidity at the time of prediction. Secondly, a key limitation of the Danish National Patient Register is the lack of precise dates of hospital diagnoses requiring us to instead use the end date of the hospitalization with the first T2D diagnosis and the start date of the subsequent hospitalization comprising the given outcome, to define the timepoints of the T2D and outcome diagnoses, respectively. Given that a fraction of our study cohort (5.6%) had their first T2D diagnosis during a relatively long hospitalization (>30 days) our approach may exclude relevant predictive events that occurred before the true date of the T2D diagnosis. Thirdly, logistic ridge regression, random forest and gradient boosting are regarded as interpretable models - *i.e.* models which predictions can be analyzed to gain an understanding of their predictions. However, the highly correlated nature of register features makes direct interpretation of feature importance challenging. This deficiency could be partially addressed by approaches such as LIME[72], Anchors[73], Shapely Additive Explanations[74] or Layerwise Relevance Propagation[75], which were designed to identify the most predictive features for each individual separately as opposed to feature importance averaged across all individuals. Analysis of patterns in individual-level feature importance could reveal non-linear relationships between features as well as patient sub-groups with different risk factor profiles. Finally, we acknowledge that future work should include additional models, extended parametrization and use other evaluation frameworks such as the one proposed employed by Pers *et al.*[76] and performance measures such as area under precision-recall curve[77].

Consequently, there is a number of additional aspects of the proposed framework that could be improved. Firstly, our approach did not leverage the temporal dimension in the data, potentially missing predictive information stored in order or time distribution of predictive events. Recently, studies which applied models incorporating time to electronic health record data suggest that inclusion of temporal information improves prediction[22,78]. Secondly, instead of providing binary prediction (i.e. predicting whether an individual will be diagnosed with an outcome or not within a certain time period) it would be advantages to perform probabilistic estimation of time-to-outcome-diagnosis. By additionally estimating a date of an adverse event, model predictions could better inform the type or intensity of an intervention. Similarly, future models could additionally estimate time and risk of death which could inform intervention considerations and simplify the learning task. Lastly, the proposed framework could be extended to leverage additional sources of information such as molecular data, lab measurements and socioeconomic information. The reference model model should then be updated to include additional variables that are more directly relevant to the outcomes studies here and typically part of the corresponding medical risk scores.

In summary, we show that inclusion of longitudinal nationwide Danish health register data comprising drug prescriptions, hospital diagnoses, hospital procedures and claims from primary care contractors moderately improves prediction of T2D comorbidities compared to a reference model based on readily-available canonical predictors. Our findings indicate that Danish health registers may be used to develop nationwide *in silico* screening approaches to identify chronic disease patients at risk of developing a certain comorbidity, but stress that, most likely, additional data - for instance genotyping, lab measurements or socio-economic data - is needed to more robustly accomplish that goal.

Data availability

The Danish health register data used in this study is available for research through Statistics Denmark[79] and the Danish Health Data Authority[80].

Received: 19 June 2019; Accepted: 16 January 2020;
Published online: 04 February 2020

References

1. World Health Organization. World report on ageing and health. *Available at*, https://www.who.int/ageing/events/world-report-2015-launch/en/ (2015).
2. Busse, R., Blümel, M., Scheller-Kreinsen, D. & Zentner, A. *Tackling chronic disease in Europe: Strategies, interventions and challenges*, vol. 20 (WHO Regional Office Europe, 2010).
3. Gaede, P., Lund, A. H., Parving, H. H. & Pedersen, O. Effect of a multifactorial intervention on mortality in type 2 diabetes. *The New Engl. journal medicine* **358**, 580–591, https://doi.org/10.1056/NEJMoa0706245 (2008).
4. Zulman, D. M., Vijan, S., Omenn, G. S. & Hayward, R. A. The relative merits of population-based and targeted prevention strategies. *The Milbank quarterly* **86**, 557–80, https://doi.org/10.1111/j.1468-0009.2008.00534.x (2008).
5. Platt, J. M., Keyes, K. M. & Galea, S. Efficiency or equity? Simulating the impact of high-risk and population intervention strategies for the prevention of disease. *SSM - Popul. Heal.* **3** (2017).
6. Jacobs-van der Bruggen, M. A. *et al*. Lifestyle interventions are cost-effective in people with different levels of diabetes risk: Results from a modeling study. *Diabetes Care* **30**, https://doi.org/10.2337/dc06-0690 (2007).
7. Chen, J. H. & Asch, S. M. Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations. *New Engl. J. Medicine* **376**, 2507–2509, https://doi.org/10.1056/NEJMp1702071 (2017).
8. Kivlahan, C. *et al*. High-Risk-Patient Identification: Strategies for Success. Tech. Rep. September, Association of American Medical Colleges, Washington, D.C. (2016).
9. Breeze, P. R. *et al*. Cost-effectiveness of population-based, community, workplace and individual policies for diabetes prevention in the UK. *Diabet. Medicine* **34**, 1136–1144, https://doi.org/10.1111/dme.13349 (2017).
10. Goldstein, B. A., Navar, A. M., Pencina, M. J. & Ioannidis, J. P. A. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J. Am. Med. Informatics Assoc.* **24**, 198–208, https://doi.org/10.1093/jamia/ocw042 (2017).
11. Saria, S., Butte, A. & Sheikh, A. Better medicine through machine learning: What's real, and what's artificial? *PLoS medicine* **15**, e1002721, https://doi.org/10.1371/journal.pmed.1002721 (2018).
12. Parikh, R. B., Kakad, M. & Bates, D. W. Integrating predictive analytics into high-valuecare the dawn of precision delivery. *JAMA - J. Am. Med. Assoc.* **315**, 651–652, https://doi.org/10.1001/jama.2015.19417 (2016).
13. Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A. & Escobar, G. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Heal. Aff.* **33**, 1123–1131, https://doi.org/10.1377/hlthaff.2014.0041 (2014).
14. Christodoulou, E. *et al*. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22, https://doi.org/10.1016/j.jclinepi.2019.02.004 (2019).
15. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE* **12**, e0174944, https://doi.org/10.1371/journal.pone.0174944 (2017).
16. Ross, E. G. *et al*. The use of machine learning for the identification of peripheral artery disease and future mortality risk. *J. Vasc. Surg*, https://doi.org/10.1016/j.jvs.2016.04.026 (2016).
17. Ye, C. *et al*. Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. *J. medical Internet research*, https://doi.org/10.2196/jmir.9268 (2018).
18. Wallert, J., Tomasoni, M., Madison, G. & Held, C. Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data. *BMC Med. Informatics Decis. Mak.*, https://doi.org/10.1186/s12911-017-0500-y (2017).
19. Arslan, A. K., Colak, C. & Sarihan, M. E. Different medical data mining approaches based prediction of ischemic stroke. *Comput. Methods Programs Biomed.*, https://doi.org/10.1016/j.cmpb.2016.03.022 (2016).
20. Unnikrishnan, P. *et al*. Development of Health Parameter Model for Risk Prediction of CVD Using SVM. *Comput. Math. Methods Medicine* **2016**, https://doi.org/10.1155/2016/3016245 (2016).
21. Kim, J. K., Kang, S. & Korea, S. Neural Network-based Coronary Heart Disease Risk Predict i on using Feature Correlation Analysis. *J. Healthc. Eng.* **2017** (2017).
22. Choi, E., Schuetz, A., Stewart, W. F. & Sun, J. Using recurrent neural network models for early detection of heart failure onset. *J. Am. Med. Informatics Assoc.* **292**, ocw112, https://doi.org/10.1093/jamia/ocw112 (2016).
23. Razavian, N. *et al*. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. *Big Data* **3**, 277–287, https://doi.org/10.1089/big.2015.0020 (2015).
24. Alghamdi, M. *et al*. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PLoS ONE* https://doi.org/10.1371/journal.pone.0179805 (2017).
25. Casanova, R. *et al*. Prediction of incident diabetes in the jackson heart study using high-dimensional machine learning. *PLoS ONE*, https://doi.org/10.1371/journal.pone.0163942 (2016).
26. Anderson, A. E. *et al*. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. *J. Biomed. Informatics*, https://doi.org/10.1016/j.jbi.2015.12.006 (2016).
27. Jahani, M. & Mahdavi, M. Comparison of predictive models for the early diagnosis of diabetes. *Healthc. Informatics Res.*, https://doi.org/10.4258/hir.2016.22.2.95 (2016).
28. Choi, B. G. *et al*. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. *Yonsei Med. J.* **60**, 191–199, https://doi.org/10.3349/ymj.2019.60.2.191 (2019).
29. Kate, R. J., Perez, R. M., Mazumdar, D., Pasupathy, K. S. & Nilakantan, V. Prediction and detection models for acute kidney injury in hospitalized older adults. *BMC Med. Informatics Decis. Mak.*, https://doi.org/10.1186/s12911-016-0277-4 (2016).

30. Corey, K. M. *et al*. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. *PLoS Medicine* 1–19, https://doi.org/10.1371/journal.pmed.1002701 (2018).
31. Ratliff, J. K. *et al*. Predicting occurrence of spine surgery complications using big data modeling of an administrative claims database. *J. Bone Jt. Surg. - Am. Vol.*, https://doi.org/10.2106/JBJS.15.00301 (2016).
32. Allyn, J. *et al*. A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: A decision curve analysis. *PLoS ONE*, https://doi.org/10.1371/journal.pone.0169772 (2017).
33. Belliveau, T. *et al*. Developing Artificial Neural Network Models to Predict Functioning One Year After Traumatic Spinal Cord Injury. *Arch. Phys. Medicine Rehabil.* https://doi.org/10.1016/j.apmr.2016.04.014 (2016).
34. Thottakkara, P. *et al*. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS ONE*, https://doi.org/10.1371/journal.pone.0155705 (2016).
35. Luo, Y. *et al*. Predicting congenital heart defects: A comparison of three data mining methods. *PLoS ONE*, https://doi.org/10.1371/journal.pone.0177811 (2017).
36. Zhang, C., Garrard, L., Keighley, J., Carlson, S. & Gajewski, B. Subgroup identification of early preterm birth (ePTB): Informing a future prospective enrichment clinical trial design. *BMC Pregnancy Childbirth*, https://doi.org/10.1186/s12884-016-1189-0 (2017).
37. Huang, S. H., Loh, J. K., Tsai, J. T., Houg, M. F. & Shi, H. Y. Predictive model for 5-year mortality after breast cancer surgery in Taiwan residents. *Chin. J. Cancer*, https://doi.org/10.1186/s40880-017-0192-9 (2017).
38. Taylor, R. A. *et al*. Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Acad. Emerg. Medicine*, https://doi.org/10.1111/acem.12876 (2016).
39. Mortazavi, B. J. *et al*. Analysis of Machine Learning Techniques for Heart Failure Readmissions. *Circ. Cardiovasc. Qual. Outcomes*, https://doi.org/10.1161/CIRCOUTCOMES.116.003039 (2016).
40. Frizzell, J. D. *et al*. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure. *JAMA Cardiol.*, https://doi.org/10.1001/jamacardio.2016.3956 (2017).
41. Mahajan, S., Burman, P. & Hogarth, M. Analyzing 30-day readmission rate for heart failure using different predictive models. In *Studies in Health Technology and Informatics*, https://doi.org/10.3233/978-1-61499-658-3-143 (2016).
42. Kulkarni, P., Smith, L. D. & Woeltje, K. F. Assessing risk of hospital readmissions for improving medical practice. *Heal. Care Manag. Sci.*, https://doi.org/10.1007/s10729-015-9323-5 (2016).
43. Sushmita, S. *et al*. Predicting 30-day risk and cost of "all-cause" hospital readmissions. *The Work. Thirtieth AAAI Conf. on Artif. Intell.* 453–461 (2016).
44. Tong, L., Erdmann, C., Daldalian, M., Li, J. & Esposito, T. Comparison of predictive modeling approaches for 30-day all-cause non-elective readmission risk. *BMC Med. Res. Methodol.*, https://doi.org/10.1186/s12874-016-0128-0 (2016).
45. Xue, Y., Liang, H., Norbury, J., Gillis, R. & Killingworth, B. Predicting the risk of acute care readmissions among rehabilitation inpatients: A machine learning approach. *J. Biomed. Informatics* **86**, 143–148, https://doi.org/10.1016/J.JBI.2018.09.009 (2018).
46. Chen, J. H., Alagappan, M., Goldstein, M. K., Asch, S. M. & Altman, R. B. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int. J. Med. Informatics* **102**, 71–79, https://doi.org/10.1016/j.ijmedinf.2017.03.006 (2017).
47. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The Parable of Google Flu: Traps in Big Data Analysis. *Sci.* **343**, 1203–1205, https://doi.org/10.1126/science.1248506 (2014).
48. Frank, L. EPIDEMIOLOGY:When an Entire Country Is a Cohort. *Sci.* **287**, 2398–2399, https://doi.org/10.1126/science.287.5462.2398 (2000).
49. Thygesen, L. C., Daasnes, C., Thaulow, I. & Brønnum-Hansen, H. Introduction to Danish (nationwide) registers on health and social issues: Structure, access, legislation, and archiving. *Scand. J. Public Heal.* **39**, 12–16, https://doi.org/10.1177/1403494811399956 (2011).
50. Schmidt, M. *et al*. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin. epidemiology* **7**, 449–90, https://doi.org/10.2147/CLEP.S91125 (2015).
51. Pottegård, A. *et al*. Data Resource Profile: The Danish National Prescription Registry. *Int. J. Epidemiol.* **46**, dyw213, https://doi.org/10.1093/ije/dyw213 (2016).
52. Andersen, J. S., De, N., Olivarius, F. & Krasnik, A. The Danish National Health Service Register. *Scand. J. Public Heal.* **39**, 34–37, https://doi.org/10.1177/1403494810394718 (2011).
53. Bliddal, M., Broe, A., Pottegård, A., Olsen, J. & Langhoff-Roos, J. The Danish Medical Birth Register. *Eur. J. Epidemiol.* **33**, 27–36, https://doi.org/10.1007/s10654-018-0356-1 (2018).
54. Helweg-Larsen, K. The Danish Register of Causes of Death. *Scand. J. Public Heal.* **39**, 26–29, https://doi.org/10.1177/1403494811399958 (2011).
55. Cessie, S. L. & Houwelingen, J. C. V. Ridge Estimators in Logistic Regression. *Appl. Stat.* **41**, 191, https://doi.org/10.2307/2347628 (1992).
56. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32, https://doi.org/10.1023/A:1010933404324 (2001).
57. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Annals Stat.* **29**, 1189–1232 (2001).
58. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. 14th international joint conference on Artif. intelligence - Vol.* **2 2**, 1137–1143 (1995).
59. Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 61–74 (MIT Press, 1999).
60. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. *ICML 2005 - Proc. 22nd Int. Conf. on Mach. Learn.* 625–632, https://doi.org/10.1145/1102351.1102430 (2005).
61. Naeini, M. P., Cooper, G. F. & Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian Binning. *Proc. Natl. Conf. on Artif. Intell.* **4**, 2901–2907 (2015).
62. Thunder, M., Moore, D. S. & McCabe, G. P. 16.2 Bootstrap t confidence intervals. In *Introduction to the Practice of Statistics* (W. H. Freeman and Company, 2007).
63. Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. & Eng.* **9**, 10–20, https://doi.org/10.1109/MCSE.2007.58 (2007).
64. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. & Eng.* **13**, 22–30, https://doi.org/10.1109/MCSE.2011.37 (2011).
65. Mckinney, W. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python for High Perform. Sci.Comput.* (2011).
66. Pedregosa, F., Weiss, R. & Brucher, M. Scikit-learn: Machine Learning in Python. *J. machine learning research* **12**, 2825–2830 (2011).
67. Smith, N. J. *et al*. Patsy: describing statistical models in Python using symbolic formulas, https://doi.org/10.5281/ZENODO.1472929 (2018).
68. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM SIGKDD Int. Conf. on Knowl. Discov. Data Min. - KDD '16* 785–794, https://doi.org/10.1145/2939672.2939785 1603.02754 (2016).
69. Bokeh Development Team. Bokeh: Python library for interactive visualization, https://bokeh.org (2019).
70. Vancheri, F., Backlund, L., Strender, L.-E., Godman, B. & Wettermark, B. Time trends in statin utilisation and coronary mortality in Western European countries. *BMJ Open* **6**, e010500, https://doi.org/10.1136/bmjopen-2015-010500 (2016).
71. Christensen, D. H., Rungby, J. & Thomsen, R. W. Nationwide trends in glucose-lowering drug use, Denmark, 1999–2014. *Clin. Epidemiol.* **8**, 381–387, https://doi.org/10.2147/CLEP.S113211 (2016).

72. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135—-1144 1602.04938 (2016).
73. Ribeiro, M. T., Singh, S. & Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)* (2018).
74. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, 4765—-4774 1705.07874 (Curran Associates, Inc., 2017).
75. Bach, S. *et al*. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, 1–46, https://doi.org/10.1371/journal.pone.0130140 (2015).
76. Pers, T. H., Albrechtsen, A., Holst, C., Sørensen, T. I. A. & Gerds, T. A. The validation and assessment of machine learning: A game of prediction from high-dimensional data. *PLoS One* **4**, https://doi.org/10.1371/journal.pone.0006287 (2009).
77. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* **10**, e0118432, https://doi.org/10.1371/journal.pone.0118432 (2015).
78. Ma, F. *et al*. Dipole. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, 1903–1911, https://doi.org/10.1145/3097983.3098088 1706.05764 (ACM Press, New York, New York, USA, 2017).
79. Data for research - Statistics Denmark. *website*, https://www.dst.dk/en/TilSalg/Forskningsservice (2019).
80. Forskerservice - Sundhedsdatastyrelsen. *website*, https://sundhedsdatastyrelsen.dk/da/forskerservice (2019).

Acknowledgements

This work was supported by a research grant from the Danish Diabetes Academy funded by the Novo Nordisk Foundation. Furthermore, THP and PD acknowledge the Novo Nordisk Foundation (grant number NNF18CC0034900) and Lundbeck Foundation (grant number R190-2014-3904 to THP). We thank Jonatan Thompson, Andreas Rieckmann and Asker Brejnrod for their comments, conducive discussions and insights.

Author contributions

P.D. and T.H.P. conceived this study. P.D. performed all the analyses. T.A.G., K.R., M.M. and H.H. participated in discussions and provided helpful methodological suggestions. P.D., T.H.P. and M.A. wrote the manuscript. All authors reviewed and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-58601-7.

Correspondence and requests for materials should be addressed to T.H.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.