

Design Matters in Patient-Level Prediction: Evaluation of a Cohort vs. Case-Control Design When Developing Predictive Models in Observational Healthcare Datasets

Jenna Reys (✉ jreps@its.jnj.com)

Janssen Research and Development Titusville <https://orcid.org/0000-0002-2970-0778>

Patrick B. Ryan

Janssen Research and Development Titusville

Peter R. Rijnbeek

Erasmus Medical Centre: Erasmus MC

Martijn J. Schuemie

Janssen Research and Development Titusville

Research

Keywords: Prediction, classification, prognostic, case-control, cohort, patient-level prediction

Posted Date: March 23rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-341457/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Journal of Big Data on August 16th, 2021. See the published version at <https://doi.org/10.1186/s40537-021-00501-2>.

Abstract

Background: The study design used to develop prediction models in observational healthcare databases (e.g., case-control and cohort) may impact the clinical usefulness. We aim to quantify how the choice of design impacts prediction model performance.

Aim: To empirically investigate differences between models developed using a case-control design and a cohort design.

Methods: Using a US claims database, we replicated two published prediction models (dementia and type 2 diabetes) which were developed using a case-control design, and also train models for the same prediction questions using cohort designs. We validated each model on data mimicking the point in time the models would be applied in clinical practice. We calculate the models' discrimination and calibration-in-the-large performances.

Results: The dementia models obtained area under the receiver operating characteristics of 0.560 and 0.897 for the case-control and cohort designs respectively. The type 2 diabetes models obtained area under the receiver operating characteristics of 0.733 and 0.727 for the case-control and cohort designs respectively. The dementia and diabetes case-control models were both poorly calibrated, whereas the dementia cohort model achieved good calibration. We show that careful construction of a case-control design can lead to comparable discriminative performance as a cohort design, but case-control designs generally oversample the outcome leading to miscalibration.

Conclusion: Any case-control design can be converted to a cohort design. We recommend that researchers with observational data use the less subjective and generally better calibrated cohort design. However, if a carefully constructed case-control design is used, then the model must be prospectively validated using a cohort design for fair evaluation and be recalibrated.

Background

It is widely known in medicine that prevention is better than cure. Prognostic models that can determine a personalized risk of some future illness could be used to identify high risk individuals that would benefit from interventions. Making decisions based on personalized risk could improve patient care [1]. Big observational healthcare databases, such as electronic health records or insurance claims data, are potential data sources that could be used to develop patient-level prediction models. A recent review of prognostic models for cardiovascular outcomes showed that the number of models being published is increasing over time, but the majority of published models have issues (e.g., methodology details missing in publication, lack of external validation and standard performance measures not used) [2]. This problem is observed across outcomes where many models fail to adhere to best practices for model development and reporting [2-4]. In addition, many published models have not been widely tested on diverse populations, so the models may perform poorly when transported to different patients such as low-income populations [4]. There may be even bigger problems with some prognostic models developed

on observational data due to study design. Prognostic models are developed by firstly creating labelled data consisting of a set of predictors for each patient and labels indicating whether each patient has the outcome during the time at risk. Machine learning algorithms are then applied to the labelled data to learn associations between the predictors and the outcome label. The idea is that these associations will generalize to new data (often in a clinical setting). It is widely known that if you provide junk data to machine learning algorithms you will get useless models returned. The study design (e.g., case-control or cohort) determines the quality of the data and therefore the quality of the model. When a model is only internally validated, the model is evaluated using the same database and study design as the model development data. If there are issues in the study design, then this is unlikely to be identified by internal validation, but the negative consequences may become apparent on external validation or when used in a real clinical setting.

The two most widely implemented study designs for extracting labelled data from observational databases are the cohort design [5] and the case-control design [6,7]. Figure 1 illustrates the differences between the designs. In Figure 1, part A shows a set of patients and their medical timelines from birth to death. Healthcare databases often only capture a section of a patient's medical observations. An index point in time is required when developing prediction models using observational data, where data prior to index are used to construct predictors and data after are used for labelling. Part B illustrates that the index for a case-control is the outcome date and part C shows that for the cohort design the index date is when a patient satisfies some specified criteria (e.g., experiences some medical event).

In the cohort design, a group of patients who you wish to be able to predict some outcome risk for, termed the 'target population', enter the cohort at a point in time where they satisfy some entrance criteria [5]. The patients are followed for some time-at-risk period to identify whether they develop the outcome. For example, to predict stroke in patients with atrial fibrillation, the target population would be 'patients newly diagnosed with atrial fibrillation' with the index of initial diagnosis, the outcome would be 'stroke' and the time-at-risk would be 1 day to 5 years following index. A patient-level prediction models is then learned by finding differences at baseline between the target population who did not experience the outcome compared to the target population who did. Alternatively, a case-control design [6,7] picks the point in time when a set of patients experience some outcome (cases), then finds some other patients (controls) paired with a date that matches on certain criteria (such as age and gender) but have no record of the outcome. The design requires the user specifying some time period prior to the outcome event where the data are used to learn to discriminate between the outcome patients and matched patients. For example, to predict stroke in patients with atrial fibrillation, the cases would be patients with stroke and a history of atrial fibrillation, and the controls would be patients with no stroke during a specified time period who have a history of atrial fibrillation and match the cases on certain criteria. The index is the stroke date for the cases and a randomly chosen date for the controls.

Case-control designs are known to have numerous issues. It is widely known that case-control designs are problematic when you wish to assess absolute risk [8]. A recent study argued that case-control designs have a temporal bias which impacts their ability to predict the future [9] and it is widely accepted

that the design leads to miscalibrated predictions. Researchers have argued that external validation of case-control prognostic models using a cohort design is essential [10]. When researchers have access to electronic health records or other longitudinal healthcare datasets, they can choose what design to use. Unfortunately, prediction models developed using the case-control design are still being published even when the researchers could have used a cohort design [11-13]. If the case-control design results in researchers extracting inappropriate labelled data, then the models developed using a case-control design may be invalid clinically even though they appear to perform well during model development (i.e., on the test set).

In this paper we empirically investigate various theoretical issues that can occur when using the case-control design to develop prediction models using observational databases. We provide examples to show the case-control design can be avoided, when a researcher has access to observational data, since any prediction problem can be properly translated into a cohort design. We replicate two published patient-level prediction studies that employed a case-control design and show that a cohort design could have been used to achieve equivalent or better discrimination and better calibration.

Issues with case-control design using observational data

Table 1 highlights that the case-control design may be problematic due to the potential issue with selection bias and lack of a well-defined point in time to apply the model. These issues can be seen in Figure 1. There are no well-defined criteria indicating when the case-control model should be applied clinically but the cohort design model is applicable when the index target population criteria is satisfied. The case-control design controls could be very healthy patients compared to the cases. In addition, the case-control design often has an incorrect matching ratio and controls are under-sampled. This is likely to impact performance metrics such as the area under the precision-recall curve and calibration and may lead to optimistic internal validation performance. It is important that a model’s predicted risks are correct when using prognostic models for decision making (i.e., if the model tells ten people they have a 10% risk, then one of them should experience the outcome). If a model overestimates risk, then interventions may be given to people unnecessarily. If a model underestimates risk, then a patient who could benefit from an intervention may be missed. Over or under-sampling outcomes often leads to models that are miscalibrated for the clinical setting they will be implemented, this is a key issue with the case-control design.

Table 1 The potential issues with the case-control designs

Issue	Description	Cohort	Case-control
Subjective data extraction methodology choices	The design requires subjective methodology choices that may differ between researchers	Not if problem is well defined with specified target population, outcome and time-at-risk	Yes – matching choice can differ (e.g., matching criteria, matching ratio, whether to remove unmatched cases)
Selection bias	Data used to train model may not be representative of target population	No	Potentially due to poor matching design
Covariate issue / protopathic bias[12]	Includes problematic covariates that are actually precursors of the outcome (e.g., symptoms/tests of outcome)	Potentially if the target cohort start date is poorly designed. Easily solved by improving target cohort criteria or adding a gap between index and time-at-risk (e.g., predict outcome 60 days to 365 days after index)	Potentially an issue if using data around outcome record (e.g., 1 day before). Can be difficult to solve.
Performance metric bias	Optimistic performance reported due to under-sampling non-outcomes	No	Potentially if matching ratio not representative of true outcome ratio (e.g., precision will be higher in data with case control ratio more equal compared to actual data)
Miscalibration issue	The predicted risk does not match the true risk	Yes (moderate chance) – if the outcome proportion changes over time or the machine learning model does not calibrate well	Yes (high chance) – if the outcome proportion is not representative due to over/under-sampling or the machine learning model does not calibrate well
Ill-defined time to apply model	No clear point in time for clinical implementation of model (where the performance has been assessed)	No – index well defined by target cohort criteria	Yes – no clear index

Defining any prediction problem as a cohort design

We assert that any prediction problem, including those previously evaluated as case-control designs, can be appropriately implemented within a cohort design. In general, a cohort design will consist of a target population (patients you want to predict the outcome for) and an index event corresponding to when you want to predict the outcome occurring. We present the different types of prediction problems and provide example inclusion criteria and index dates for defining the problem as a cohort design, see Table 2.

Table 2 Different types of prediction problems and examples of how they fit the cohort design

Prediction type	Target population	outcome	Example target cohort inclusion criteria and index	
			Target cohort criteria	Index
Disease onset	General population	Disease (e.g., depression)	A visit (outpatient or inpatient) during 2010, >365 days observation in database, age ≥ 18 , no prior illness	First valid visit in 2010
Disease progression	Early stage disease patients	Advanced stage disease	Diagnosed with disease, >365 days observation in database	Initial disease record date
Treatment choice	Patients dispensed treatment 1 or 2	Treatment 1	Dispensed treatment 1 or 2, >365 days observation in database	First recorded date of treatment 1 or 2
Treatment response	Patients dispensed a treatment	Desired effect (e.g., disease cured)	Dispensed treatment at adequate therapeutic level, >365 days observation in database	First recorded date of treatment
Treatment safety	Patients dispensed a treatment	An adverse event	Dispensed treatment, >365 days observation in database	First recorded date of treatment
Treatment adherence	Patients dispensed a treatment	>X% days covered during some follow-up	Dispensed treatment, >365 days observation in database	First recorded date of treatment

Methods

Replication of case-control patient-level predictions and cohort comparison

We selected two published patient-level prediction models that used a case-control design to develop the models using observational healthcare data. The first predicted future Alzheimer's risk⁶ and the second predicted future type 2 diabetes risk [7]. We replicated the two case-control models by following the published process, but because we do not have access to the same patient-level data, we instead use the

Optum® De-Identified Clinformatics® Data Mart Database – Socio-Economic Status (Optum Claims), a US claims database. Optum Claims contains outpatient pharmacy dispensing claims, inpatient and outpatient medical claims which provide procedure codes and diagnosis codes. The data also contain selected laboratory test results for a non-random sample of the population. We used data prior to December 31 2014 to develop the Alzheimer's model and data prior to November 30 2012 to develop the type 2 diabetes model, to best match the data used in the published papers. We also developed equivalent cohort design models for both outcomes where the target population was patients with a healthcare visit and the outcomes were the same as those used in the development of the case-control.

The use of Optum Claims was reviewed by the New England Institutional Review Board (IRB) and were determined to be exempt from broad IRB approval.

Models were developed using a Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression model trained on 75% of the data and internally validated on the remaining 25% of the data. The optimal hyper-parameter (regularization value) was determined using 3-fold cross validation on the 75% data used to train the model.

To fairly compare the performance of the two designs we applied the models when patients visit their healthcare provider and have not experienced the outcome before, but used data collected in a time period after the data used to develop the models. This was accomplished by applying the models to predict the 3-year risk of Alzheimer's and type 2 diabetes at the point in time a patient visits their healthcare provider and is free of the disease being predicted. For this evaluation we used an 'external' validation set: visits post December 31 2014 for Alzheimer's and post November 30 2012 for type 2 diabetes. This validation aims to mimic how the models would be used clinically.

Replication study 1: Dementia:

Case-control data construction

Following the design by Albrecht et al [6], cases were defined as patients age 18 or older diagnosed with dementia or prescribed a dementia drug for the first time between 2008-01-01 and 2014-12-31, with 1095 days prior observation and 180 days post observation. Patients must also have another record of dementia or dementia drug in the 180 days following. Patients were excluded if they had a diagnosis for nutritional deficiencies or alcohol or substance dependency within 3 months of the index date, or had a hospice claim during the 6 months previous to the index. Controls were selected based on matching on age, gender and having a visit within 30 days around the case index date but were excluded if they had a dementia drug or condition record, had a diagnosis for nutritional deficiencies or alcohol or substance dependency within 3 months of the index date, or had a hospice claim during the 6 months previous to the index. 4 controls were matched per case. The index date was the date of the initial dementia record for the cases and the matching visit date for the controls.

Candidate predictors were constructed using conditions, procedures, measurements, observations and visit counts recorded between 1095 days prior and 730 days prior to index. We also included age at index and gender variables.

Cohort data construction

To reformulate the prediction problem as a cohort design, we defined the target population as patients with visit between 2008-01-01 and 2011-12-31 who were aged 18 or older with no prior dementia conditions or drug records and 365 days or more prior observation. Index date is first valid visit. The outcome was the first record of dementia condition or drug with another dementia condition or drug recorded in the 180 days following with no diagnosis for nutritional deficiencies or alcohol or substance dependency within 3 months prior and no hospice claim during the 6 months prior. For those in the target population we predicted who will have the outcome 1 day after index until 1095 days after index (within 3 years after index). There were a large number of patients (12,861,202) with a valid visit, so we sampled 1,000,000 patients from the target population for model development.

For consistency between designs, we used similar predictors constructed using conditions, procedures, measurements, observations and visit counts recorded between 365 days prior and 0 days prior to index. We also included age at index and gender variables.

Validation data construction

To evaluate how a model would perform in a realistic clinical setting, we picked a validation population consisting of eligible target patients: the first visit a patient had between 2015-01-01 and 2015-12-31 satisfying a minimum of 365 days observation prior to index, aged 18 or older and no prior dementia condition or drug records. The outcome was the same as defined for the cohort data construction and we predicted whether the outcome would occur 1 day after index until 1095 days after index (within 3 years after index).

We then applied the models generated from the case-control and cohort designs for each patient at their first valid visit and evaluated the models' performances in predicting the 3-year risk of dementia.

Replication study 2: Type 2 diabetes:

Following the design by McCoy et al. [7], we defined cases as patients aged 18 to 89 diagnosed with type 2 diabetes or prescribed a type 2 diabetes drug for the first time between 2008-01-01 and 2012-11-30, with 1095 days prior observation. Patients must also have another record of type 2 diabetes condition or drug in the 180 days following the initial event. Patients were excluded if any of the following criteria were met:

- they had a record of disorders of pancreatic internal secretion (ICD-9 code 251.8) within 1095 days prior to the index date
- they had a record of poisoning by adrenal cortical steroids (ICD-9 code 962.0) within 1095 days prior to the index date

- they had a record of secondary diabetes diagnosis (ICD-9 codes 249.x) any time prior

Controls were selected based on matching on location and enrolment time. 10 controls were matched per case without replacement from a pool of candidate eligible controls who were aged 18 to 89 and:

- No type 2 diabetes condition or drug recorded prior to 2012-11-30
- No record of disorders of pancreatic internal secretion (ICD-9 code 251.8) within 1095 days prior to the index date
- No record of poisoning by adrenal cortical steroids (ICD-9 code 962.0) within 1095 days prior to the index date
- No record of secondary diabetes diagnosis (ICD-9 codes 249.x) any time prior

The index date was the date of the first record of type 2 diabetes condition or drug for cases and the matched case's index date for the controls. The authors also stated that they excluded patients with only routine care records or no encounters [7]. As it was not clear how this would be defined in our data, we chose to remove patients with less than 3 condition records to ensure cases and controls were active in the databases.

Candidate predictors were constructed using conditions and drugs recorded between 1095 days prior and 1 days prior to index. We also included age at index, gender, ethnicity and race variables.

Cohort data construction

Reformulated as a cohort design, the target population were defined to be patients with a visit between 2008-01-01 and 2009-11-30 who were aged between 18 and 89 with no prior type 2 diabetes conditions or drug records and 365 days or more prior observation. Patients were excluded if they had disorders of pancreatic internal secretion or poisoning by adrenal cortical steroids in the prior 365 days or secondary diabetes diagnosis any time prior to index. Index date is first valid visit. The outcome was the first record of type 2 diabetes condition or drug with another type 2 diabetes condition or drug recorded in the 180 days following. For those in the target population we predicted who will have the outcome 1 day after index until 1095 days after index (within 3 years after index). There were 7,966,573 patients with a valid visit, so we sampled 4,000,000 patients from the target population for model development. This sample size was chosen so that case-control and cohort designs had a similar number of outcomes.

For consistency between designs, we used similar predictors constructed using conditions, and drugs recorded between 365 days prior and 0 days prior to index plus age at index, gender, ethnicity and race.

Validation data construction

To evaluate the models in a realistic clinical setting we picked a validation population consisting of eligible target patients: the first visit a patient had between 2012-12-01 and 2014-12-31 satisfying a minimum of 365 days observation prior to index, aged between 18 and 89, no prior type 2 diabetes condition or drug records, no disorders of pancreatic internal secretion or poisoning by adrenal cortical

steroids in the prior 365 days and no secondary diabetes diagnosis any time prior to index. The outcome was the same as defined for the cohort data construction and we predicted whether the outcome would occur 1 day after index until 1095 days after index (within 3 years after index).

We then apply the models from the case-control and cohort designs for each patient at their first valid visit and evaluated the models' performances in predicting the 3-year risk of type 2 diabetes.

Results

Dementia

For the case control we identified 118,694 eligible cases in Optum Claims. Restricting the cases to those with 4 or more matching controls left 11,016 cases and 44,064 controls. We excluded 8,671 ineligible controls (met exclusion criteria) to end up with a final dataset containing 11,016 cases and 35,393 controls in Optum claims. The case-control model was trained using the 46,409 patients with 11,016 dementia patients (~ 250 in 1000) and obtained an internal area under the receiver operating characteristic curve (AUC) of 0.657. This was consistent with the original development paper's reported performance of 0.65. The cohort model was trained using a 1,000,000 target patient sample with 4,108 patients (~ 4 in 1000) diagnosed with dementia within 3 years of their visit. The cohort model obtained an internal AUC of 0.944 (0.937–0.950).

Inspecting the models showed that the case control model lacked the age variables that were included in the cohort model. This is due to the cases being matched to controls on age and gender, so neither of these will be predictive in the case-control design. Many other variables seemed to be included in both models (including amnesia, organic mental disorder, Parkinson's disease, mood disorder, seizure and memory impairment).

The validation data contained 12,264,784 patients with 103,518 (~ 8 in 1000) having dementia recorded within 3 years. The case-control model obtained an AUC of 0.560 and the cohort model obtained an AUC of 0.897. The discrimination difference was due to age not being included in the case control model. When evaluating both models on subsets of patients within each five-year age group the cohort model discrimination performance was still better than the case control model, see Table 3. The mean observed dementia risk in the validation data was 0.84%. The cohort model's mean predicted risk was 0.70%, indicating the cohort model slightly under-estimated risk. The case-control model's mean predicted risk was 23.95%, so it severely over-estimated risk. The discrimination and calibration plots can be seen in Fig. 2.

Table 3
The AUC performance of both models predicting dementia when applied to each subset of patients within each 5-year age group

Dementia Temporal Validation AUC performance		
Age group	Cohort design	Case control design
ALL	0.897	0.560
age group: 18–19	0.652	0.511
age group: 20–24	0.668	0.547
age group: 25–29	0.683	0.520
age group: 30–34	0.684	0.595
age group: 35–39	0.673	0.572
age group: 40–44	0.699	0.604
age group: 45–49	0.699	0.601
age group: 50–54	0.712	0.629
age group: 55–59	0.726	0.653
age group: 60–64	0.720	0.671
age group: 65–69	0.689	0.661
age group: 70–74	0.666	0.642
age group: 75–79	0.650	0.634
age group: 80–84	0.631	0.629
age group: 85–89	0.613	0.616

Type 2 Diabetes

For the case-control model we found 65,991 eligible patients aged 18 to 89 diagnosed with type 2 diabetes or prescribed a type 2 diabetes drug for the first time between 2008-01-01 and 2012-11-30, with 1095 days prior observation. 213 patients were excluded due to pancreatic internal secretion, 3 were excluded due to poisoning by adrenal cortical steroids and 93 were excluded due to secondary diabetes. A further 952 were excluded due to having less than 3 condition records between 2008-01-01 and 2012-11-30. This left us with 64,730 cases. We identified 5,974,383 patients aged 18–89 with no diabetes records prior to 2012-11-30 but with 3 or more condition records during 2008-01-01 and 2012-11-30. We excluded 2007 patients with pancreatic internal secretion, 283 patients with poisoning by adrenal cortical steroids and 0 patients with secondary diabetes. This left us with 5,972,093 candidate controls. We then matched on location and enrolment time to find 646,539 controls. The case-control model trained using

the case-control data, 711,269 patients with 64,730 patients having type 2 diabetes (~ 91 in 1000), obtained an internal AUC of 0.833. The cohort model was trained using a 3,993,438 target patient sample (4,000,000 were sampled but 6,562 of these left the database at index so had no time-at-risk and were excluded) with 54,898 patients (~ 14 in 1000) diagnosed with type 2 diabetes within 3 years of their visit. The cohort model obtained an internal AUC of 0.742.

The models appeared to contain similar predictors such as 'polycystic ovaries', 'abnormal glucose tolerance in mother complicating pregnancy', 'metabolic syndrome X', 'older age', 'hypoglycemic disorder', 'polyuria', 'chronic nonalcoholic liver disease', being 'Hispanic' and 'obesity'. The case control model identified 'glycosuria', which may be a symptom of existing type 2 diabetes rather than a predictor of future risk.

The validation data contained 8,939,289 patients with 251,659 (~ 28 in 1000) having type 2 diabetes recorded within 3 years. The case-control model obtained an AUC of 0.733 and the cohort model obtained an AUC of 0.727. When evaluating the case-control and cohort models on the subset of patients in each 5-year age group, the models performed similarly in terms of discrimination, see Table 4. The mean observed diabetes risk was 2.8%. The cohort model mean predicted risk was 1.6% (under-estimated risk) and the case-control mean predicted risk was 7.7% (over-estimating risk). The discrimination and calibration plots can be seen in Fig. 3.

Table 4

The AUC performance of both models predicting type 2 diabetes when applied to each subset of patients within each 5-year age group

Type 2 Diabetes Temporal Validation AUC performance		
Age group	Cohort design	Case control design
ALL	0.727	0.733
age group: 18–19	0.701	0.712
age group: 20–24	0.713	0.710
age group: 25–29	0.707	0.704
age group: 30–34	0.701	0.705
age group: 35–39	0.709	0.716
age group: 40–44	0.710	0.718
age group: 45–49	0.708	0.715
age group: 50–54	0.705	0.707
age group: 55–59	0.693	0.696
age group: 60–64	0.676	0.678
age group: 65–69	0.623	0.628
age group: 70–74	0.597	0.601
age group: 75–79	0.572	0.574
age group: 80–84	0.552	0.549
age group: 85–89	0.549	0.551

Discussion

This study illustrated i) that the case-control design has many potential flaws when using observational healthcare data to develop prediction models and ii) that any prediction problem trained using a case-control design can be transformed into a cohort design. We empirically compared two published case-control design prediction models against equivalent cohort designs to show that a cohort design could have been used to obtain similarly discriminative models. However, the cohort design model is more likely to be trained in a population representing the true target population and be better calibrated compared to the case-control design that often under-samples patients without the outcome.

For the cohort and case-control designs the mean predicted risk tended to be similar to the outcome proportion in the training data. This was problematic for the case-control design as the outcome is often

over sampled (or the non-outcomes under-sampled) resulting in models that drastically over-estimated risk. For example, the case-control dementia design used a 1:4 match ratio, which resulted in ~ 25% of the patients in the training data having the outcome. This caused calibration issues as only 0.84% of the target population patients had the outcome, so the case-control model design resulted in very inaccurate risk estimates. Unless the case-control matching uses the true ratio (this is highly unlikely), any models developed using this design will be miscalibrated and will require an extra recalibration step to ensure risk estimates are accurate. The cohort design was not immune to miscalibration, as the outcome proportion can change over time. This is the reason we saw slight to moderate under-estimation of risk. This is a known issue and can be reduced by restricting the data used to develop the model to more recent data [14]. The temporal change in outcome proportion observed in this study may have been inflated by the data converting to ICD-10 from ICD-9 between the model development and validation dates due to improved diabetes coding. For example, type 2 diabetes patients may have had unspecified diabetes recorded in ICD-9 but when the coding became more granular, they may have had specific type 2 diabetes recorded in ICD-10.

The case-control design and cohort design models appeared to include similar variables. However, the case-control design model sometimes included variables that appear to be symptoms/early tests of the outcome. These are not useful if the model's purpose is to predict new outcomes in patients who are outcome free. The inclusion of these variables did not seem to impact the performance in the type 2 diabetes model. The case-control design model is also unable to include variables that are used to match cases and controls. The dementia model matched on age and gender, but this resulted in these variables being missing from the model, which greatly impacted the overall discrimination. The authors⁶ recommended developing separate case-control design models for different age groups, but this strategy reduces the outcome count used for training each model and may not be possible for rare outcomes.

The case-control design model internal discriminative performance appears to be an overestimate of the true discrimination (dementia internal AUC of 0.66 but external AUC of 0.56 and type 2 diabetes internal AUC of 0.83 but external AUC of 0.72). Metrics that vary based on how unbalanced the data are, such as the precision, would be affected even more when the data becomes more imbalanced. Therefore, any model developed using a case-control design needs to be fairly evaluated on cohort design data.

The results show that a well specified case-control design can avoid selection bias issues and we did not see discriminative issues with applying the model at a random visit, even though the case-control design has an ill-defined application time. The calibration of each case-control model was poor and requires recalibration. The case-control matching is subjective, and a poorly designed matching strategy could in theory limit the generalizability of the model. In addition, the case-control designed model needs to be validated using cohort design data to fairly evaluate its performance. We recommend that a cohort design is used instead as it reduces i) subjectively, ii) the potential for selection bias, iii) miscalibration and iv) reporting of optimistic performance. We showed that a cohort design can be used for any prediction and is less prone to the issues presented in Tables 1 and 5, as long as recent data are used.

Table 5
Summary of issues observed and potential solutions

Issue	Issues observed in study		Solution
	Cohort	Case-control	
Subjective methodology choices	No	Yes – the case-control designs used different matching criteria	Use a cohort design
Selection bias	NA	Did not appear to be a problem in the two predictions investigated	NA
Covariate issue	NA	1. Symptoms appeared in the diabetes model but didn't impact performance. 2. The dementia model was unable to include variables used to match controls	Use covariates to stratify patients and develop separate models
Performance metric bias	Yes – due to temporal changes the internal validation was slightly optimistic	Yes – due to incorrect matching ratios and potentially non-generalizable development population the internal validation was very optimistic	Perform external validation with cohort design to fairly assess performance
Miscalibration	Some due to temporal changes the risk was under-estimated	Yes – due to incorrect matching ratios the risk was over-estimated in both examples	Train models on more recent data Recalibrate if necessary
Ill-defined time to apply model	NA	Not a problem for the two predictions investigated – the models appeared to perform reasonably when applied at the validation index event (even though they were not developed using this index)	NA

Conclusions

In this paper we discussed potential issues when developing a prediction model from labelled data constructed using a case-control design from electronic healthcare or claims data. We argued that using a cohort design for developing prediction models is preferred and overcomes bias and clinical application issues that can plague the case-control design. We replicated two published prediction models developed using a case-control design and showed that these models could have been developed with a cohort design. The cohort design models had equivalent discrimination compared to case-control design models when applied to data representing realistic clinical applications of the models. However, the cohort design models were better calibrated than the case-control design models. Calibration is important, as accurate individual risk estimates are needed when using models clinically for decision making. The AUC discrimination metric only provides a measure of how well a model can rank patients based on risk. A

highly discriminative model could be harmful for decision making if it is not well calibrated. The case-control design is more difficult to implement since it requires the specification of often subjective matching criteria. This may have a big impact on the model's generalizability. As a result, we recommend that other researchers either avoid using a case-control design when developing patient-level prediction models using observational healthcare data or ensure they validate any case-control design model on cohort design data and perform any recalibration if necessary. The cohort design ensures a well-defined point in time for applying the model, provides fairer performance metrics and results in a better calibrated model.

Abbreviations

IRB - Institutional Review Board

LASSO - Least Absolute Shrinkage and Selection Operator

ICD - International Classification of Diseases

AUC – Area under the receiver operating characteristic curve

Declarations

Ethics approval and consent to participate

The use of Optum Claims was reviewed by the New England Institutional Review Board (IRB) and were determined to be exempt from broad IRB approval.

Consent for publication

Not Applicable

Availability of data and materials

The Optum claims data that support the findings of this study are available from Optum (contact at: <http://www.optum.com/solutions/data-analytics/data/real-world-data-analytics-a-cpl/claims-data.html>) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Competing interests

JMR, PBR and MJS are employees of Janssen R&D and shareholders of Johnson and Johnson.

Funding

PRR received funding from the European Health Data and Evidence Network (EHDEN) project of the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Authors' contributions

All authors contributed to the study design. JMR executed the analysis. All authors read and approved the final manuscript.

Acknowledgements

Not Applicable

References

1. Croft P, Altman DG, Deeks JJ, Dunn KM, Hay AD, Hemingway H, LeResche L, Peat G, Perel P, Petersen SE, Riley RD. The science of clinical practice: disease diagnosis or patient prognosis? Evidence about “what is likely to happen” should shape clinical practice. *BMC Med*. 2015;13(1):20.
2. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiochia V, Roberts C, Schlüssel MM. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:p.i2416.
3. Moons KG, Altman DG, Reitsma JB, Collins GS. New guideline for the reporting of studies developing, validating, or updating a multivariable clinical prediction model: the TRIPOD statement. *Adv Anat Pathol*. 2015;22(5):303–5.
4. Haniffa R, Isaam I, De Silva AP, Dondorp AM, De Keizer NF. Performance of critical care prognostic scoring systems in low and middle-income countries: a systematic review. *Crit Care*. 2018;22(1):18.
5. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc*. 2018;25(8):969–75.
6. Albrecht JS, Hanna M, Kim D, Perfetto EM. Predicting diagnosis of Alzheimer's disease and related dementias using administrative claims. *Journal of managed care & specialty pharmacy*. 2018;1138–1145.
7. McCoy RG, Nori VS, Smith SA, Hane CA. Development and validation of HealthImpact: an incident diabetes prediction model based on administrative data. *Health services research*. 2016;51(5):1896–918.
8. Steyerberg EW. Clinical prediction models. Vol. 381. New York: Springer; 2009.
9. Yuan W, Beaulieu-Jones BK, Yu KH, Lipnick SL, Palmer N, Loscalzo J, Cai T, Kohane IS. Temporal bias in case-control design: preventing reliable predictions of the future. *Nat Commun*. 2021;12(1):1–10.

10. Ten Haaf K, Steyerberg EW. Methods for individualized assessment of absolute risk in case-control studies should be weighted carefully. *Eur J Epidemiol.* 2016;31(11):1067–8.
11. Chien LH, Chen CH, Chen TY, Chang GC, Tsai YH, Hsiao CF, Chen KY, Su WC, Wang WC, Huang MS, Chen YM. Predicting lung cancer occurrence in never-smoking females in Asia: TNSF-SQ, a prediction model. *Cancer Epidemiology and Prevention Biomarkers.* 2020 Feb 1;29(2):452-9.
12. Mandair D, Tiwari P, Simon S, Rosenberg M. DEVELOPMENT OF A PREDICTION MODEL FOR INCIDENT. MYOCARDIAL INFARCTION USING MACHINE LEARNING APPLIED TO HARMONIZED ELECTRONIC HEALTH RECORD DATA. *Journal of the American College of Cardiology.* 2020 Mar 24;75(11_Supplement_1):194-.
13. Ho WK, Tan MM, Mavaddat N, Tai MC, Mariapun S, Li J, Ho PJ, Dennis J, Tyrer JP, Bolla MK, Michailidou K. European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. *Nature communications.* 2020 Jul;31(1):1–1. 11(.
14. Sharon E, Davis TA, Lasko G, Chen ED, Siew ME, Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc.* 2017;24(6):1052–61.

Figures

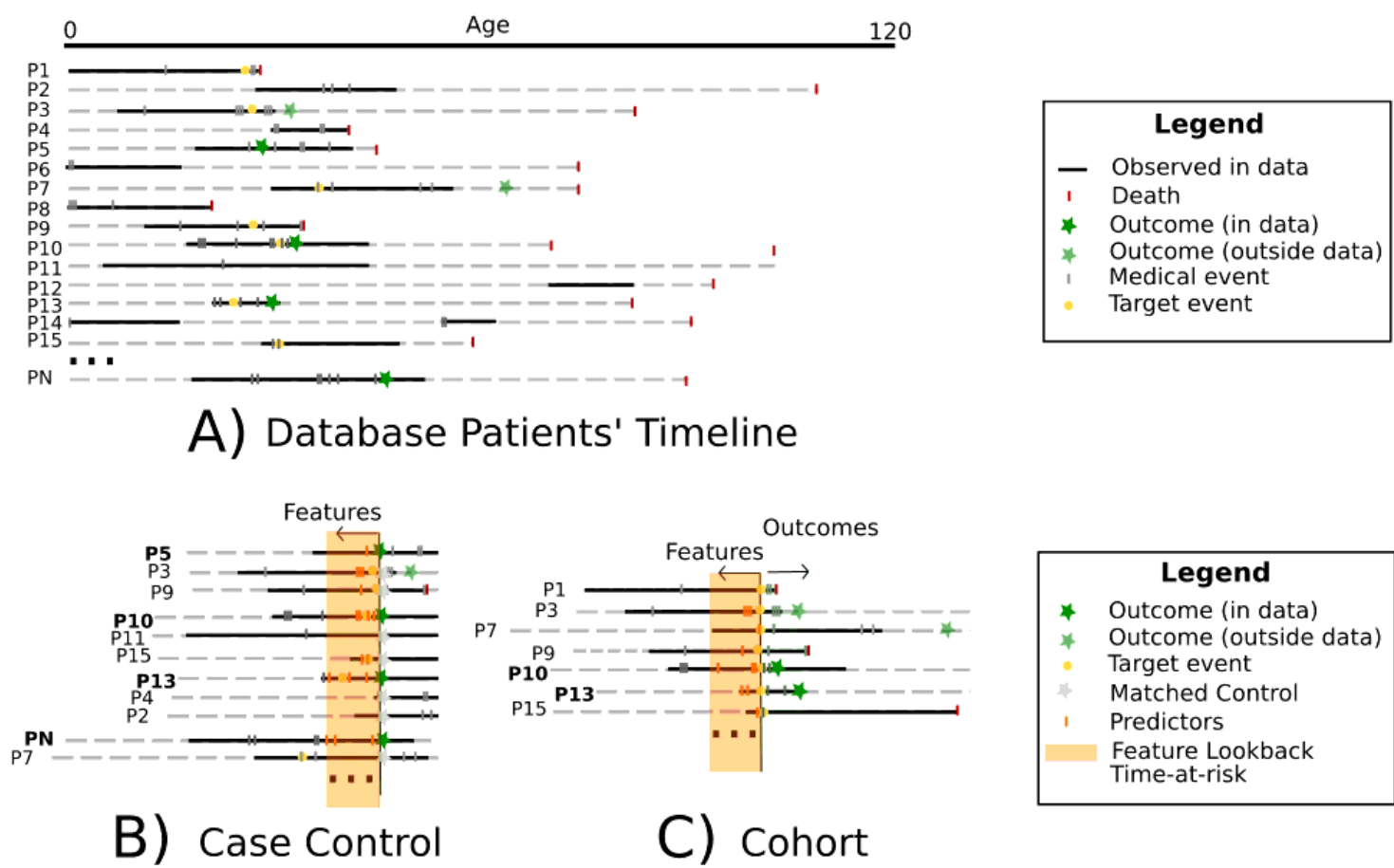


Figure 1

Illustration of how longitudinal data (part A) can be converted into labelled data using the case-control design (part B) and the cohort design (part C). The case-control design centers around the outcome date and uses data prior to this date to construct features. The cohort design centers around the data a patient satisfies criteria, using data prior to this to construct features and data post this to identify outcome labels.

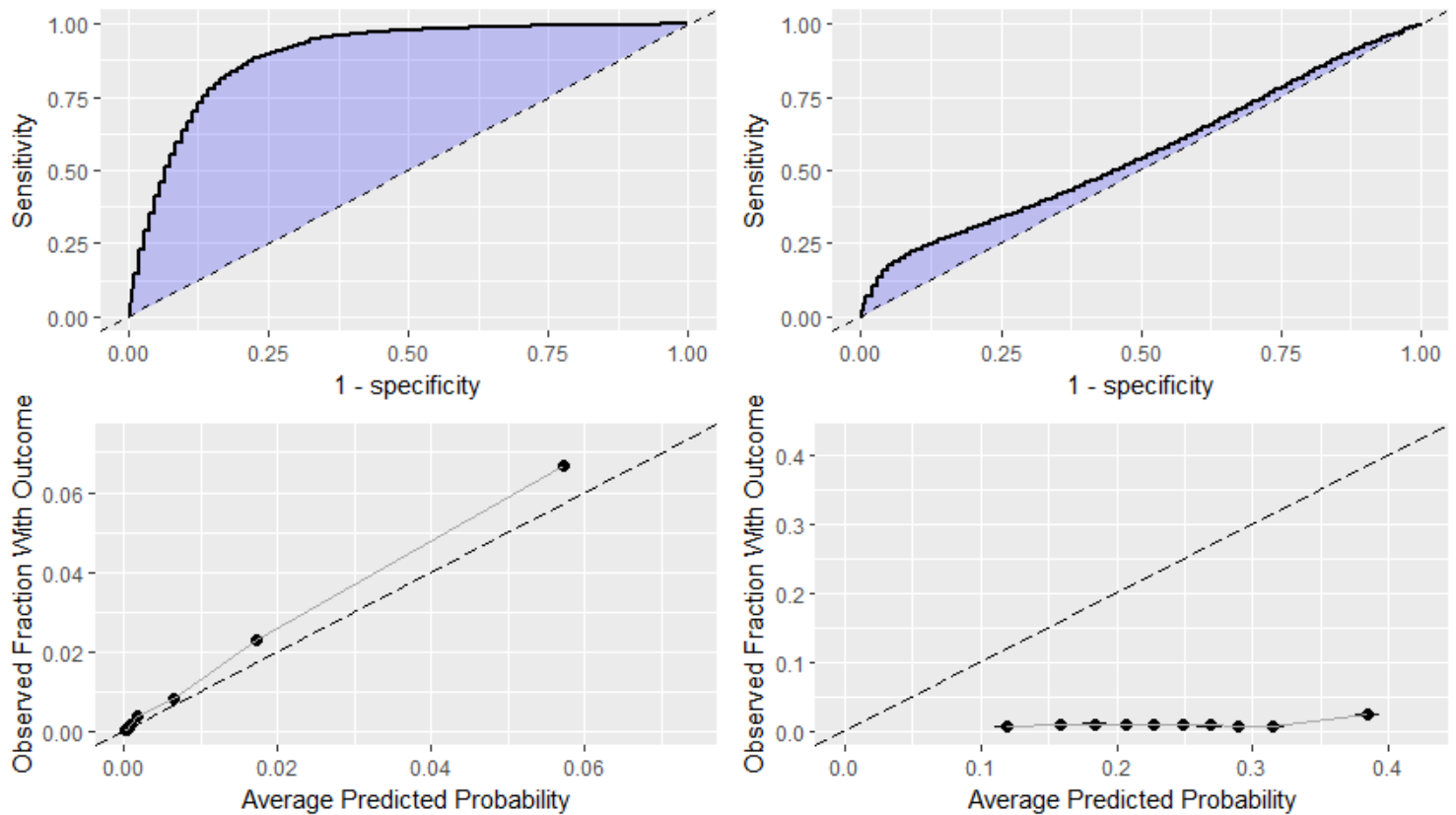


Figure 2

The ROC and calibration plots for the models predicting dementia on the validation dataset (cohort left and case control right)

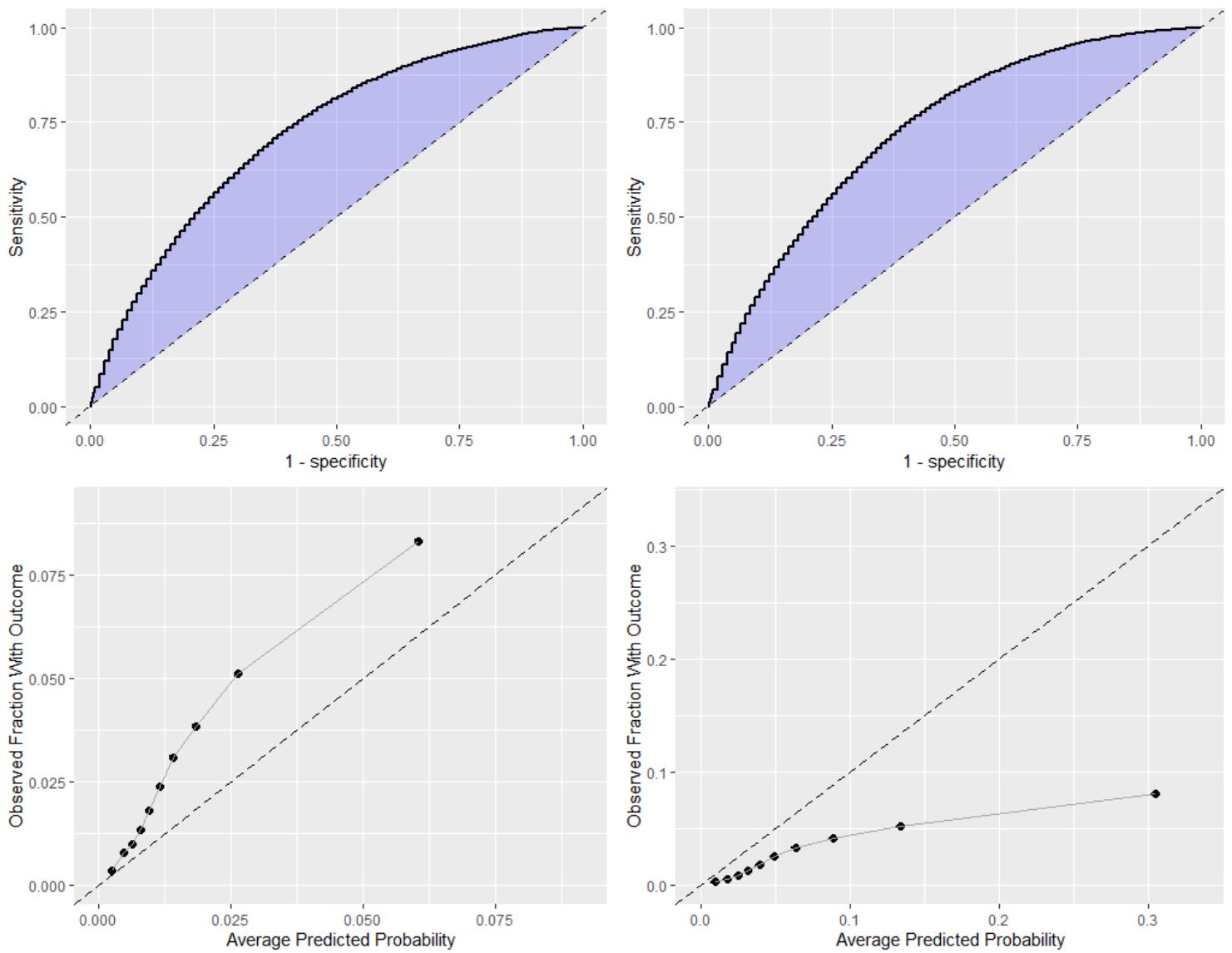


Figure 3

ROC and calibration plots for predicting type 2 diabetes models on the validation dataset (cohort left and case control right)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixImages.docx](#)