BMJ Open

Recording type 2 diabetes mellitus in a standardised central Saudi database: a retrospective validation study

Turki Abdulaziz Althunian ,[1,2] Meshael M Alrasheed,[3] Fatemah A Alnofal,[1] Rawan T Tafish,[4] Mahmood A Mira,[4] Raseel A Alroba,[1] Mohammed W Kirdas,[4] Thamir M Alshammari [5]

To cite: Althunian TA, Alrasheed MM, Alnofal FA, *et al*. Recording type 2 diabetes mellitus in a standardised central Saudi database: a retrospective validation study. *BMJ Open* 2023;**13**:e065468. doi:10.1136/bmjopen-2022-065468

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-065468).

Received 13 June 2022
Accepted 19 February 2023

[1]Research Informatics Department, Saudi Food and Drug Authority, Riyadh, Saudi Arabia
[2]College of Medicine, Alfaisal University, Riyadh, Saudi Arabia
[3]Executive Department for Research and Studies, Saudi Food and Drug Authority, Riyadh, Saudi Arabia
[4]Orthopedic and Spinal Surgery, Kingdom Hospital & Consulting Clinics, Riyadh, Saudi Arabia
[5]College of Pharmacy, Almaarefa Univeristy, Riyadh, Saudi Arabia

Correspondence to
Dr Thamir M Alshammari; thamer.alshammary@gmail.com

ABSTRACT

Objectives This study was conducted to assess the validity of recording (and the original diagnostic practice) of type 2 diabetes mellitus at a hospital whose records were integrated to a centralised database (the standardised common data model (CDM) of the Saudi National Pharmacoepidemiologic Database (NPED)).

Design A retrospective single-centre validation study.

Settings Data of the study participants were extracted from the CDM of the NPED (only records of one tertiary care hospital were integrated at the time of the study) between 1 January 2013 and 1 July 2018.

Participants A random sample of patients with type 2 diabetes mellitus (≥18 years old and with a code of type 2 diabetes mellitus) matched with a control group (patients without diabetes) based on age and sex.

Outcome measures The standardised coding of type 2 diabetes in the CDM was validated by comparing the presence of diabetes in the CDM versus the original electronic records at the hospital, the recording in paper-based medical records, and the physician re-assessment of diabetes in the included cases and controls, respectively. Sensitivity, specificity, positive predictive value and negative predictive value were estimated for each pairwise comparison using RStudio V.1.4.1103.

Results A total of 437 random sample of patients with type 2 diabetes mellitus was identified and matched with 437 controls. Only 190 of 437 (43.0%) had paper-based medical records. All estimates were above 90% except for sensitivity and specificity of CDM versus paper-based records (54%; 95% CI 47% to 61% and 68%; 95% CI 62% to 73%, respectively).

Conclusions This study provided an assessment to the extent of which only type 2 diabetes mellitus code can be used to identify patients with this disease at a Saudi centralised database. A future multi-centre study would help adding more emphasis to the study findings.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ We examined the validity of using only the code of type 2 diabetes mellitus in cohort and outcome identification versus three reference standards in the standardised electronic health records (EHRs) of a single hospital.
⇒ We further examined the validity of two additional algorithms to identify type 2 diabetes in cohort and outcome identification in the original EHRs of the hospital.
⇒ To our knowledge, our study was the first of its kind in the region.
⇒ Our study was limited by including only one centre.
⇒ Not including all type 2 diabetes mellitus-related diagnostic codes was another limitation.

INTRODUCTION

Data collected electronically from the provision of routine clinical care (ie, real-world data (RWD)) have been used to generate evidence (real-world evidence (RWE)) on benefits, risks and the usage of pharmaceuticals.[1–10] In Saudi Arabia, the electronic recording of health data in hospital settings has increased at the major tertiary hospitals during the last decade.[11–14] In 2018, the Saudi Food and Drug Authority established the National Pharmacoepidemiologic Database (NPED) to integrate and standardise electronic health records (EHRs) from different hospitals in Saudi Arabia.[11] The NPED was initiated to maximise the usage of RWE in supporting drug regulatory decision-making processes.[11] The NPED will also be used in determining disease natural histories and trends in Saudi Arabia.[11] A standardisation was performed for the EHRs that were imported from the first hospital using the Observational Health Data Sciences and Informatics common data model (CDM).[11] The standardisation process was followed by an initial data quality assessment (no alarming concerns were identified). However, this quality assessment did not include assessing the validity of the recorded data.[11] Additionally, and up to our knowledge, no study has been published to assess the validity of the health recording practice at any of the Saudi hospitals (especially those of disease diagnostic codes).

The validity of RWD is integral in conducting pharmacoepidemiological

research studies.[8 15 16] Conducting validation studies in the Saudi healthcare system would assist not only in improving the quality of the generated RWE but also in supporting stakeholders in implementing their quality improvement initiatives. Validating the diagnostic codes of diabetes (especially type 2 diabetes mellitus) is a priority given its high prevalence in Saudi Arabia (up to 25% of the Saudi population was estimated to have diabetes with an increased prevalence of 51% among the 70 to 79-year-old population), and given the lack of well-designed and large-scale pharmacoepidemiological studies in the Saudi population with diabetes.[17–19] Studies have shown that the validity of recording diabetes mellitus in the context of RWD has been assessed in different health records using different types of data sources (eg, physician claims, hospital discharge data, EHRs), with different reference standards (mostly medical records, self-reported or telephone surveys) and different case definitions (eg, using one diagnostic code or one claim for diabetes mellitus (and/or another indicator of diabetes mellitus such as high glucose levels), two or more codes/claims).[20–24]

This study was conducted to assess the validity of the original, the extracted and the standardised diagnostic codes of type 2 diabetes mellitus of the EHRs that were imported from the first hospital to the NPED. The validity of the original diagnosis of type 2 diabetes mellitus at that hospital was also assessed. Finally, the study was aimed to assess whether the diagnostic code of type 2 diabetes can be used to identify patients with type 2 diabetes (or diabetes as an outcome) in the standardised EHRs of that hospital.

METHODS

Study design, data source and patient population

This study was a retrospective single-centre validation study. The study was carried out using the EHRs that were imported and mapped from a 129-bed private tertiary care hospital in Riyadh (the imported EHRs included a record of at least 500 000 patients) to the NPED. A sample of patients with type 2 diabetes mellitus (one code of type 2 diabetes mellitus), who visited the hospital in the period between 1 January 2013 and 1 July 2018, was randomly selected from the standardised EHRs of the hospital (ie, CDM), then (using the standardised EHRs) a control group (patients without type 2 diabetes mellitus) was randomly matched based on age and sex (a control group was included for the estimation of specificities and negative predictive values). The included participants were required to be ≥18 years and have at least one health record.

Validation methods

The standardised diagnostic code of diabetes in the CDM was validated using a three-step validation approach. The first validation step was aimed to confirm the presence (in the included cases) and the absence (in the included controls) of type 2 diabetes in the sample that was extracted from the CDM by reviewing the patients' original EHRs at the hospital (the first reference). Diseases were coded during the study period at the hospital using the International Statistical Classification of Diseases and Related Health Problems 9th Revision (ICD-9) code (ICD-9 code of type 2 diabetes mellitus: 250.00). This validation step will help in assessing the accuracy of the CDM standardisation process at the NPED, the completeness of the EHR extraction process, and the validity of the original coding ICD coding of type 2 diabetes mellitus at the hospital.

Box 1 Criteria for diagnosing type 2 diabetes mellitus

- ⇒ FPG ≥126 mg/dL (7.0 mmol/L).*
- ⇒ 2-h PG ≥200 mg/dL (11.1 mmol/L) during an OGTT.
- ⇒ Haemoglobin A1c ≥6.5% (48 mmol/mol).
- ⇒ Symptoms of hyperglycaemia or hyperglycaemic crisis (polyuria, polydipsia and unexplained weight loss), AND a random plasma glucose ≥200 mg/dL (11.1 mmol/L).
- ⇒ On therapy for diabetes mellitus (antidiabetic medications) and previous diagnosis of diabetes mellitus in medical records.

FPG, fasting olasma glucose; 2-h PG, 2-h plasma glucose; OGTT, oral glucose tolerance test.
*In the absence of unequivocal hyperglycaemia, these criteria should be confirmed by repeat testing on a different day.

The second validation assessment was performed using a different reference: the patients' paper-based medical records. This validation step also included a comparison between the original EHRs (the first reference) versus paper-based medical records (the second reference) as an additional step to validate the former reference. In the final validation step, which was also a step to validate the original diagnosis of type 2 diabetes mellitus at the hospital, all study patients (both cases and controls) were re-assessed for the presence of type 2 diabetes by one of the study physicians (the third reference) based on the hospital criteria which are adopted from the American Diabetes Association classification and diagnosis of diabetes (box 1).[25] The physician was allowed to use all resources at the hospital that are necessary to complete the diagnosis. The code of patients with type 2 diabetes in the standardised CDM was compared with the third reference. The findings from the assessment of the third reference were also compared with those of the first (original EHRs) and the second (paper-based medical records) references as an additional validation step of the latter two references. An additional step to confirm the diagnosis of type 2 diabetes mellitus in the original EHRs was performed using two algorithms:

- First algorithm: type 2 diabetes code+a prescription of an antidiabetic medication.
- Second algorithm: type 2 diabetes code+a prescription of an antidiabetic medication+a blood measurement reflective of diabetes.

Table 1 Validation estimates of the pairwise comparisons among all study patients

Comparisons	Validation estimates (%)			
	PPV (95% CI)	Sn (95% CI)	NPV (95% CI)	Sp (95% CI)
CDM vs EHRs (n=874)	1.00 (0.99 to 1.00)	0.93 (0.90 to 0.95)	0.92 (0.89 to 0.95)	1.00 (0.99 to 1.00)
CDM vs paper-based records	0.54 (0.47 to 0.61)	0.93 (0.86 to 0.97)	0.96 (0.92 to 0.98)	0.68 (0.62 to 0.73)
CDM vs re-diagnosis (n=874)	1.00 (0.99 to 1.00)	0.93 (0.90 to 0.95)	0.92 (0.89 to 0.95)	1.00 (0.99 to 1.00)
EHRs vs re-diagnosis (n=806*)	1.00 (0.99 to 1.00)	1.00 (0.99 to 1.00)	1.00 (0.99 to 1.00)	1.00 (0.99 to 1.00)
EHRs vs paper-based records (n=336*)	0.53 (0.45 to 0.61)	0.98 (0.92 to 1.00)	0.99 (0.96 to 1.00)	0.68 (0.62 to 0.74)

*Controls in the CDM that were identified as cases in EHRs were excluded from this analysis with their corresponding cases.
CDM, common data model; EHR, electronic health record; NPV, negative predictive value; PPV, positive predictive value; Sn, sensitivity; Sp, specificity.

These algorithms were chosen based on clinical judgement and based on other previously published algorithms.[20 21]

The degree to which the results of the algorithm assessments agree with the code-only analysis will also provide more information on whether only the code can be used to identify this population in both the CDM and EHRs. Sensitivity, specificity, the positive predictive value (PPV) and the negative predictive value (NPV) were the targeted parameters in this study. The values of these validation estimates might give an indication of the extent to which only the diagnostic code of type 2 diabetes mellitus can be used to identify type 2 diabetes mellitus as an outcome in the CDM.

Statistical analysis

The validation parameters were estimated for each pairwise comparison with their corresponding 95% CIs. To demonstrate a sensitivity of 85% and an expected width of 95% CI of 10% (taking into account the 25% prevalence of type 2 diabetes mellitus in Saudi Arabia), a minimum sample size of 196 patients (98 cases and 98 controls) was needed.[17 18 26] The minimum total sample sizes for validating the first and second algorithms in the original EHRs were 138 and 73, respectively. All statistical analyses were performed using RStudio V.1.4.1103.

Patient and public involvement

Patients and the public were not involved in the design, conduct or reporting/dissemination of this study.

RESULTS

Table 1 shows the number of patients who were included in each pairwise comparison. A total of 437 random patients with type 2 diabetes mellitus (427 (98.0%) were on antidiabetics and/or had haemoglobin A1c (HbA1c) measurement of >6.5%) were identified and matched with 437 controls. Almost one-third of the cases were identified before 2016 (141 of 437 (32.3%)). Of the totally matched pairs, only 190 (43.0%) had paper-based medical records. The median age of the included patients (both the cases and controls) was 56 years (IQR=21), and 522 of the included patients (60.0%) were men. Type 2 diabetes mellitus (among the cases) was diagnosed between 2007 and 2018. The majority of the cases (83.6%) had abnormal HbA1c levels at the time of (or

Table 2 Validation estimates of the assessed algorithms

Comparisons	Validation estimates (%)			
	PPV (95% CI)	Sn (95% CI)	NPV (95% CI)	Sp (95% CI)
EHRs vs re-diagnosis (n=784) Code and antidiabetic(s) (1st algorithm)	0.97 (0.95 to 0.98)	1.00 (0.99 to 1.00)	1.00 (0.99 to 1.00)	0.97 (0.95 to 0.98)
EHRs vs re-diagnosis (n=598) Code and antidiabetic(s) and HbA1c >6.5% (2nd algorithm)	1.00 (0.99 to 1.00)	1.00 (0.99 to 1.00)	1.00 (0.98 to 1.00)	1.00 (0.98 to 1.00)

EHR, electronic health record; HbA1c, haemoglobin A1c; NPV, negative predictive value; PPV, positive predictive value; Sn, sensitivity; Sp, specificity.

within 6 months) the index date (the date of diagnosing type 2 diabetes mellitus).

The estimates of validating the standardised code of type 2 diabetes mellitus in the CDM versus two references (EHRs and the re-diagnosis) were all above 90% (table 1). The validation estimates for EHRs versus re-diagnosis were also above 90%. The PPV and specificity for CDM versus paper-based documentation were 46% and 32% lower compared with those versus EHRs and re-diagnosis (table 1). Of 190 cases that were included in the validation assessment with the paper-based documentation as a reference, 87 (46%) did not have any records for type 2 diabetes mellitus in their medical charts. Type 2 diabetes mellitus among these 87 was mostly diagnosed after 2013 (the year of the large-scale usage of EHRs at the hospital), and only 8 patients were diagnosed with diabetes before 2013. This may justify the absence of diabetes recording in their paper-based medical charts. The results of the validation assessment of the first and second algorithms in the EHRs were comparable with that of the type 2 diabetes mellitus code-only analysis (table 2).

DISCUSSION

This study assessed an approach to the population of type 2 diabetes mellitus using the disease-only code in standardised EHRs of a Saudi hospital. With the exception of the assessment versus paper-based records, all validation estimates of the standardised and the original codes of type 2 diabetes mellitus were above 90% (the estimates of the algorithms were almost comparable with these estimates). These findings might be supportive of using only the standardised diagnostic code to identify type 2 diabetes mellitus as an outcome in these records.

In our assessment of the validity of the standardised code of type diabetes mellitus in the CDM, the minimum value of sensitivity (93% vs paper-based records) was higher than the average minimum value observed in the previous diabetes-case definition validation studies (26.9%).[20 21 23–25] On the other hand, the minimum value of specificity (68%) was lower compared with the average minimum value observed in the published studies (88%).[20–24] The minimum values of PPV and NPV were almost comparable with the average minimum values that were observed in the published studies (54% vs 54% and 92% vs 90.8%, respectively).[20–24] Two of the references in our study (EHRs and re-diagnosis) were used as references in the previous diabetes-case validation studies.[20–24] In 44.4% (8 of 18) of the published diabetes-case definition validation studies, the original EHRs were used as a reference (the other references were the physician re-diagnosis, self-reported or telephone surveys, and a multisource approach).[20–23] Type 2 diabetes mellitus was confirmed in 100% of the cases in our study, which is almost comparable to the confirmation results in a previous study in which the re-diagnosis was used as a reference.[24] The value of validation estimates in our study might be supportive of using only the diagnostic code to identify patients with type 2 diabetes mellitus as cohorts or to identify diabetes as an outcome in the standardised EHRs that were imported from the first hospital.

Our study was the first diabetes-case definition validation study (and the first validation study for a diagnostic code) in the region. Three reference standards were used in our study, and validity was assessed at different three levels (code extraction, code standardisation and the original diagnosis of diabetes) and was compared with those of algorithms. Our study has two limitations. First, the study was a single-centre study. The generalisability may improve by conducting a multi-centre study that takes the variability of hospital coding systems into account. Second, ICD-9 was used to code type 2 diabetes mellitus at the hospital during the study period; however, the hospital (and other hospitals) started upgrading their coding system to ICD-10. Additionally, we did not include other type 2 diabetes-related ICD-9 codes (codes for uncontrolled diabetes and diabetic complications) in our assessment. Including the same hospital in a future single or multi-centre with an updated sample of ICD-10 codes of type 2 diabetes mellitus and/or its complications would help in adding more emphasis to the study findings.

We assessed the validity of the (standardised) diagnostic code of type 2 diabetes mellitus at different recording levels and provided an indication to the extent to which this code can be used to identify this disease as an outcome. A future multi-centre study that includes an updated sample from the hospital with ICD-10 codes of type 2 diabetes mellitus and/or its complications would help in adding more emphasis to the study findings.

Acknowledgements We thank the information technology (IT) and the medical records (MR) departments in the Kingdom Hospital for the help in accessing the patient records and retrieving the data required for this research.

Contributors TAA, MWK and TMA designed the study. MMA, RTT, RAA and FAA contributed to the data collection process. All authors were involved in designing, analysing and conducting the study. MAM was involved in confirming the diagnosis. TAA drafted the report. All authors critically reviewed the manuscript. TAA and TMA accept full responsibility for the work and conduct of the study, had full access to the data and controlled the decision to publish. TMA is responsible for the overall content as the guarantor.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The study was approved by the SFDA ethics committee (ethics approval number: 2020_012).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iDs
Turki Abdulaziz Althunian http://orcid.org/0000-0002-8030-7963
Thamir M Alshammari http://orcid.org/0000-0002-5630-2468

REFERENCES

1 Barda N, Dagan N, Ben-Shlomo Y, *et al*. Safety of the BNT162b2 mRNA covid-19 vaccine in a nationwide setting. *N Engl J Med* 2021;385:1078–90.
2 Hernán MA. Methods of public health research-strengthening causal inference from observational data. *N Engl J Med* 2021;385:1345–8.
3 Li M, Chen S, Lai Y, *et al*. Integrating real-world evidence in the regulatory decision-making process: a systematic analysis of experiences in the US, EU, and China using a logic model. *Front Med (Lausanne)* 2021;8:669509.
4 Beaulieu-Jones BK, Finlayson SG, Yuan W, *et al*. Examining the use of real-world evidence in the regulatory process. *Clin Pharmacol Ther* 2020;107:843–52.
5 Wu J, Wang C, Toh S, *et al*. Use of real-world evidence in regulatory decisions for rare diseases in the United States-current status and future directions. *Pharmacoepidemiol Drug Saf* 2020;29:1213–8.
6 Franklin JM, Glynn RJ, Martin D, *et al*. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clin Pharmacol Ther* 2019;105:867–77.
7 Danaei G, García Rodríguez LA, Cantero OF, *et al*. Electronic medical records can be used to emulate target trials of sustained treatment strategies. *J Clin Epidemiol* 2018;96:12–22.
8 Toh S. Pharmacoepidemiology in the era of real-world evidence. *Curr Epidemiol Rep* 2017;4:262–5.
9 Sherman RE, Anderson SA, Dal Pan GJ, *et al*. Real-world evidence-what is it and what can it tell us? *N Engl J Med* 2016;375:2293–7.
10 Dreyer NA. Making observational studies count: shaping the future of comparative effectiveness research. *Epidemiology* 2011;22:295–7.
11 Alnofal FA, Alrwisan AA, Alshammari TM. Real-world data in Saudi Arabia: current situation and challenges for regulatory decision-making. *Pharmacoepidemiol Drug Saf* 2020;29:1303–6.
12 Samra H, Li A, Soh B, *et al*. Utilisation of hospital information systems for medical research in Saudi Arabia: a mixed-method exploration of the views of healthcare and it professionals involved in hospital database management systems. *Health Inf Manag* 2020;49:117–26.
13 Alsulame K, Khalifa M, Househ M. EHealth in Saudi Arabia: current trends, challenges and recommendations. *Stud Health Technol Inform* 2015;213:233–6.
14 El Mahalli A. Adoption and barriers to adoption of electronic health records by nurses in three governmental hospitals in eastern Province, Saudi Arabia. *Perspect Health Inf Manag* 2015;12:1f.
15 Herrett E, Thomas SL, Schoonen WM, *et al*. Validation and validity of diagnoses in the general practice research database: a systematic review. *Br J Clin Pharmacol* 2010;69:4–14.
16 Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the general practice research database: a systematic review. *Br J Gen Pract* 2010;60:e128–36.
17 Robert AA, Al Dawish MA. The worrying trend of diabetes mellitus in Saudi Arabia: an urgent call to action. *Curr Diabetes Rev* 2020;16:204–10.
18 Saeedi P, Petersohn I, Salpea P, *et al*. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International diabetes Federation diabetes atlas, 9th edition. *Diabetes Res Clin Pract* 2019;157:107843.
19 Ministry of Health. World health survey Saudi Arabia: 2019 final report; Available: www.moh.gov.sa/en/Ministry/Statistics/Population-Health-Indicators/Documents/World-Health-Survey-Saudi-Arabia.pdf
20 Chen G, Khan N, Walker R, *et al*. Validating ICD coding algorithms for diabetes mellitus from administrative data. *Diabetes Res Clin Pract* 2010;89:189–95.
21 Lipscombe LL, Hwee J, Webster L, *et al*. Identifying diabetes cases from administrative data: a population-based validation study. *BMC Health Serv Res* 2018;18:316.
22 Khokhar B, Jette N, Metcalfe A, *et al*. Systematic review of validated case definitions for diabetes in ICD-9-coded and ICD-10-coded data in adult populations. *BMJ Open* 2016;6:e009952.
23 Moreno-Iribas C, Sayon-Orea C, Delfrade J, *et al*. Validity of type 2 diabetes diagnosis in a population-based electronic health record database. *BMC Med Inform Decis Mak* 2017;17:34.
24 de Burgos-Lunar C, Salinero-Fort MA, Cárdenas-Valladolid J, *et al*. Validation of diabetes mellitus and hypertension diagnosis in computerized medical records in primary health care. *BMC Med Res Methodol* 2011;11:146.
25 American Diabetes Association. 2. classification and diagnosis of diabetes: standards of medical care in diabetes-2020. *Diabetes Care* 2020;43:S14–31.
26 Buderer NM. Statistical methodology: I. incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Acad Emerg Med* 1996;3:895–900.