

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

Data Resource Profile: A National Linked Routine-Care Data Resource for Longitudinal Health Services Research in Thailand: The National Health Security Office–Type 2 Diabetes Mellitus (NHSO–T2DM) Extract, 2017–2024

Phongthana Pasookhush^{1,2}, Apinya Surawit^{1,2}, Anupong Sirirungreung³, Ploypun Narindrarangkura³, Ram Rangsin⁴, Sophida Suta¹, Sureeporn Pumeiam¹, Pichanun Mongkolsucharitkul¹, Bonggochpass Pinsawas¹, Suphawan Ophakas¹, and Korapat Mayurasakorn^{1,2,*}

Submission History

Submitted:	27/03/2026
Accepted:	20/08/2026
Published:	20/10/2026

¹Siriraj Population Health and Nutrition Research Group, Research Department, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

²Siriraj Center of Research Excellence for Diabetes and Obesity (SiCORE-DO), Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

³Department of Military and Community Medicine, Phramongkutklao College of Medicine, Bangkok, Thailand

⁴Phramongkutklao College of Medicine, Bangkok, Thailand

Abstract

Introduction

The National Health Security Office–type 2 diabetes mellitus (NHSO–T2DM) resource was created to support national longitudinal research among people with recorded incident T2DM under Thailand's Universal Health Coverage scheme during 2017–2024.

Methods

Six routine-care files—T2DM registry, non-communicable disease screening, chronic follow-up, laboratory follow-up, service encounters, and drug dispensing—were curated and linked. Files #1–#5 were linked deterministically using an encrypted person identifier (CID). Drug records in File #6 were linked to service encounters through the transaction identifier (TRAN_ID) to inherit CID and service date (DATE_SERV). Encounter records were deduplicated and collapsed to CID + DATE_SERV; laboratory results were cleaned and reshaped; and drug records were standardised and aggregated to person-day–drug exposures.

Results

The resource contains 1,421,470 individuals and three analysis-ready tables: T2DM_registry, with 1,421,470 person-level records; Master_NHSO–T2DM, with 51,160,751 person-day visit-events; and Drug_FINAL, with 62,229,519 person-day–drug exposure records. All drug TRAN_IDs matched a service record, enabling 91,626,041 drug rows to inherit CID and DATE_SERV before aggregation. Data domains include demographics, diagnoses and selected outcome dates, screening and follow-up measures, laboratory results, service costs, and dispensed medications. Quality-control outputs document processing counts, key uniqueness, linkage completeness, and missingness.

Conclusion

NHSO–T2DM provides a governed, analysis-ready resource for reproducible studies of diabetes care patterns, medication exposure, longitudinal outcomes, comparative effectiveness and safety, and risk prediction. Access is available through approved collaboration, subject to ethical and institutional governance.

Key Features

- The National Health Security Office–type 2 diabetes mellitus (NHSO–T2DM) resource is a national scale, linked routine-care dataset covering Thailand's Universal Health Coverage (UHC) system, curated into standardised analysis-ready tables to support reproducible longitudinal research.
- The NHSO–T2DM was established to enable population-level studies of type 2 diabetes care, outcomes, treatment patterns, and risk prediction using routinely collected health information.
- The dataset includes 1,421,470 individuals with recorded incident type 2 diabetes during 2017–2024 in the person-level registry, with diagnoses most commonly occurring at ages 50–69 years and representation across adult age groups.
- Person-level linkage uses an encrypted identifier (CID), and dispensed medication items are linked to service encounters using a transaction identifier (TRAN_ID) to inherit CID and service date (DATE_SERV), producing three analysis-ready master tables: T2DM_registry (person-level), Master_NHSO–T2DM (person-day visit-event), and Drug_FINAL (person-day–drug exposure).
- The resource includes multi-domain data comprising registry-defined diagnosis and selected diabetes-related outcomes with dates, screening and chronic follow-up measures, laboratory test records, service encounter diagnoses and costs, and dispensed medication items.
- The dataset is not publicly available; access is through collaboration and subject to ethical approval and institutional governance. Interested researchers should contact the corresponding author to discuss feasibility and the appropriate application pathway.

Keywords

type 2 diabetes mellitus, routine health data, universal health coverage, thailand

*Corresponding Author:

Email Address: korapat.may@mahidol.ac.th (Korapat Mayurasakorn)

Background

Thailand has developed a national routine health information system to support monitoring and reporting of key health indicators. This system is supported by investments in data management, connectivity, large-scale data storage at provincial and national levels, and cloud-based processing to enable indicator computation and reporting through national health data platforms. Within this ecosystem, a widely used institution-based data standard is commonly referred to as the “43 folders/43 files” system, supported by the Ministry of Public Health. The Ministry publishes and periodically updates operational manuals describing the standard file structure and provides accompanying standard code sets (e.g., diagnosis and procedure codes and medication-related standards) to support consistent data submission across healthcare facilities [1]. Diagnoses in encounter data are recorded using ICD-10 (International Classification of Diseases, 10th Revision) codes. Records in the 43-file standard are typically generated from facility hospital information systems and are used within national health data platforms such as the Health Data Center (HDC) for routine indicator reporting and data completeness monitoring [2]. In addition, the National Health Security Office (NHSO) provides an online channel for submission of individual-level outpatient and prevention/promotion (OP/PP) records in the 43-file format for administrative purposes [3]. However, these routinely collected records are distributed across multiple domain-specific files and cannot be used directly for longitudinal research without linkage, harmonisation, deduplication, and explicit quality-control procedures.

To address the need for a research-ready longitudinal dataset, the National Health Security Office–type 2 diabetes mellitus (NHSO–T2DM) resource was established to enable population-scale longitudinal research on type 2 diabetes mellitus (T2DM) in Thailand using routinely collected data under the Universal Health Coverage (UHC) scheme administered by the NHSO. The original research questions motivating creation of this curated extract were to: (i) quantify the incidence and timing of diabetes-related complications following recorded T2DM diagnosis; (ii) evaluate real-world medication use patterns and the comparative effectiveness of commonly used glucose-lowering therapies; and (iii) develop and validate prediction models for incident T2DM complications using routinely collected clinical measurements, laboratory test results, diagnoses, and medication dispensing. The extract comprises UHC beneficiaries with incident T2DM recorded during 2017–2024, with data captured at the level of individuals (person-level registry) and repeated routine-care encounters (visit-events), including laboratory records and dispensed medication items. The resource’s principal contribution is a documented curation framework that transforms national routine-care data into linked, analysis-ready person-level, visit-event, and medication exposure tables for reproducible diabetes and health services research.

This curated extract and its curation pipeline were established in 2025 and are maintained within a secure research computing environment at the Siriraj Population Health and Nutrition Research Group, Faculty of Medicine Siriraj Hospital, Mahidol University, and the Department of Military and Community Medicine, Phramongkutklao College of Medicine,

Bangkok, Thailand. Data were delivered by the NHSO as six comma-separated value (CSV) files and linked within the extract using an encrypted personal identifier (CID), with an encounter-level transaction identifier (TRAN_ID) linking service encounters to dispensed drug items. No linkages to external datasets were performed for the current release. The current version is a one-time frozen extract covering 2017–2024; follow-up is limited to this observation window, and no routine refresh is currently planned.

Methods

Original Sources of Data and Study Population

The NHSO–T2DM resource comprises secondary, routinely collected administrative and routine-care health information system data under Thailand’s Universal Health Coverage (UHC) scheme. The study population includes UHC beneficiaries with incident type 2 diabetes mellitus (T2DM) recorded during 2017–2024 in the person-level registry. Longitudinal follow-up is passive and event-driven, with records captured whenever individuals access care during the observation window.

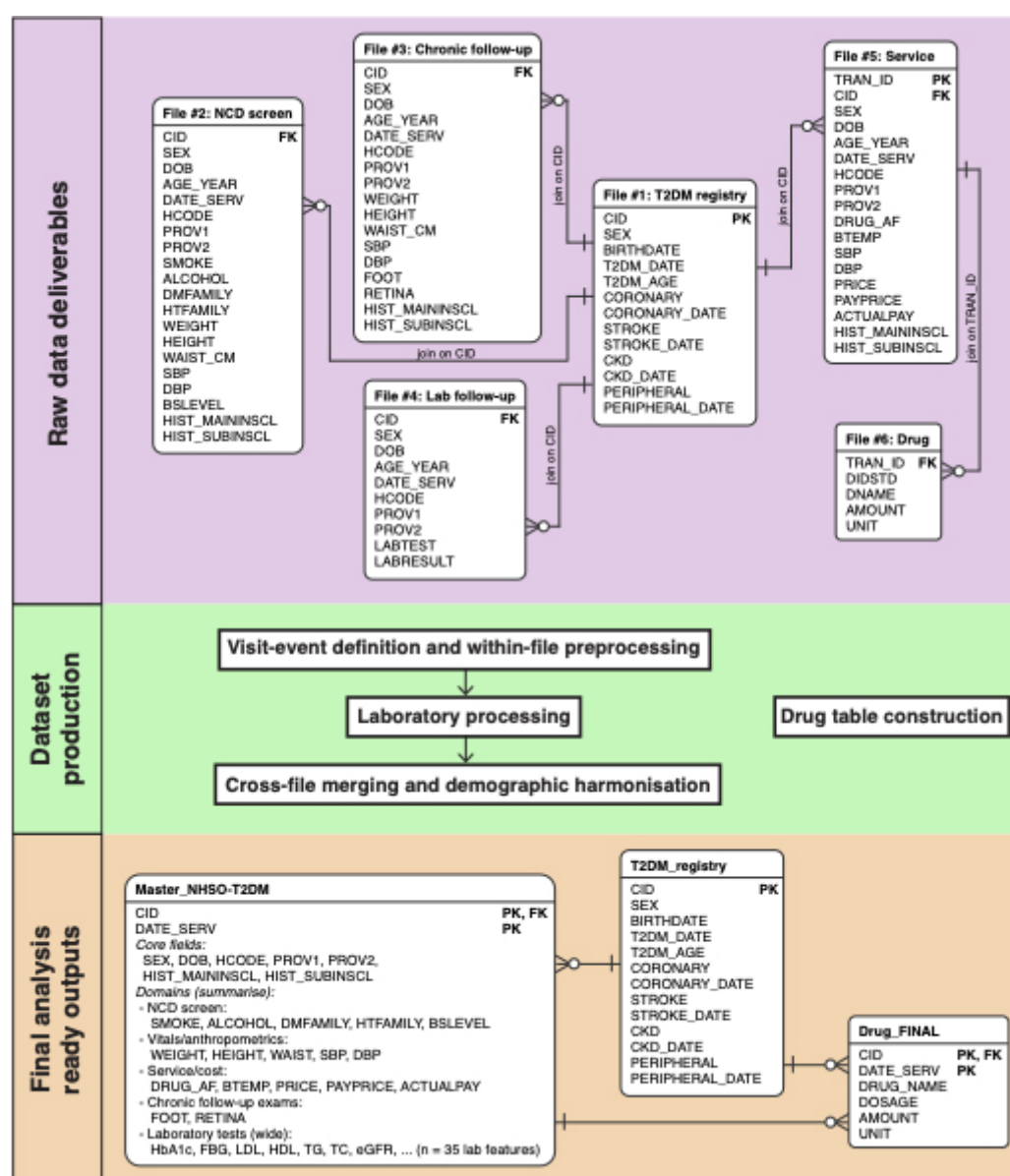
The extract was delivered by the National Health Security Office (NHSO) as six linked comma-separated value (CSV) files representing complementary domains of routine-care: (i) a person-level T2DM registry; (ii) outpatient screening records; (iii) chronic follow-up records; (iv) laboratory follow-up records; (v) service encounter records containing diagnoses and costs; and (vi) dispensed medication items. An overview of file grain, record counts, linkage fields, and example variables is provided in Table 1 and Figure 1.

Top panel shows the six delivered raw files linked by an encrypted person identifier (CID) across Files #1–#5 and by a transaction identifier (TRAN_ID) between service encounters (File #5) and dispensed drug items (File #6). Middle panel summarises key processing steps used to create analysis-ready datasets, including visit-event definition, laboratory processing, demographic harmonisation, and drug table construction. Bottom panel shows the three analysis-ready master tables: a person-level registry table (T2DM_registry), a person-day visit-event table (Master_NHSO–T2DM), and a person-day-drug table (Drug_FINAL).

Data Linkage Methods

Linkage within the extract was deterministic and used exact matching; no probabilistic or fuzzy linkage methods were applied. The encrypted personal identifier (CID), present in Files #1–#5, was used to identify records belonging to the same individual across the registry and encounter-derived domains. Within encounter-aligned Files #2–#5, records were aligned at the visit-event level using the composite key CID + DATE_SERV, where DATE_SERV represents the calendar date of service. This enabled records from different clinical domains for the same person and date to be combined before applying the prespecified deduplication and field-specific collapsing rules. Because the drug dispensing file (File #6) did not contain CID or DATE_SERV, its records were linked to service encounters in

Figure 1: Structure of the National Health Security Office–Type 2 Diabetes Mellitus (NHSO-T2DM) Data Resource and Dataset Production Pipeline



File #5 by exact matching on the encounter-level transaction identifier (TRAN_ID). Matched drug records inherited CID and DATE_SERV from the corresponding service record and were subsequently standardised and aggregated to the Drug_FINAL grain; TRAN_ID was then removed from the final table. Linkage quality was assessed by comparing distinct TRAN_ID values across Files #5 and #6 and by calculating the proportion of drug records that successfully inherited CID and DATE_SERV. No external datasets were linked in the current release.

Data Processing and Creation of Analysis-Ready Master Tables

To enable reproducible analyses across projects, the six delivered files (Table 1) were transformed into three analysis-ready master tables (Table 2): a person-level registry table (T2DM_registry), a person-day visit-event

table (Master_NHSO-T2DM), and a person-day-drug exposure table (Drug_FINAL). The processing pipeline removed exact duplicate records, reconciled multiple records representing the same person-date event, and harmonised repeated demographic and facility fields across files while preserving information content wherever possible [4]. Detailed variable-level aggregation rules, demographic consensus logic, laboratory dictionaries and parsing rules, and drug mapping and normalisation procedures are provided in the Supplementary Data.

Visit-Event Construction and Within-File Preprocessing (Files #2–#5)

Within each of Files #2–#5, exact duplicate rows were removed and remaining records sharing the same person-date were collapsed to one row per CID + DATE_SERV using predefined, field-specific aggregation and precedence rules. When repeated values occurred for shared facility, province,

Table 1: Delivered raw files in the National Health Security Office–type 2 diabetes mellitus (NHSO-T2DM) 2017–2024 extract

File	Grain (unit of record)	Rows	Linkage fields	Main domains	Example variables
File #1: T2DM registry	Person-level	1,421,470	CID	Cohort entry and diabetes-related diseases/outcomes grouped into four disease categories (Supplementary Table S5)	CID, SEX, BIRTHDATE, AGE_SERV, T2DM_DATE, CORONARY, CORONARY_DATE, STROKE, STROKE_DATE, CKD, CKD_DATE, PERIPHERAL, PERIPHERAL_DATE,
File #2: NCD screen	Person-visit	8,133,103	CID, DATE_SERV	Screening visits, lifestyle risk factors, vitals/anthropometrics, benefit categories	SMOKE, ALCOHOL, DMFAMILY, HTFAMILY, WEIGHT, HEIGHT, WAIST_CM, SBP, DBP, BSLEVEL, HIST_MAININSCL, HIST_SUBINSCL, HCODE, PROV1, PROV2
File #3: Chronic follow-up	Person-visit	32,029,769	CID, DATE_SERV	Chronic disease follow-up visits, vitals/anthropometrics, diabetes screening/exams	WEIGHT, HEIGHT, WAIST_CM, SBP, DBP, FOOT, RETINA, HIST_MAININSCL, HIST_SUBINSCL, HCODE, PROV1, PROV2
File #4: Lab follow-up	Person-visit-lab test	104,589,043	CID, DATE_SERV, LABTEST	Laboratory testing (long format)	LABTEST, LABRESULT, DATE_SERV, HCODE, PROV1, PROV2
File #5: Service	Person-visit (+transaction)	24,678,715	CID, DATE_SERV (+TRAN_ID)	Encounters/utilisation, costs, selected vitals, drug-use flag	TRAN_ID, DRUG_AF, BTEMP, SBP, DBP, PRICE, PAYPRICE, ACTUALPAY, HIST_MAININSCL, HIST_SUBINSCL, PROV1, PROV2
File #6: Drug	Transaction-drug item	91,626,041	TRAN_ID	Dispensed drug items (code/name/amount/unit)	TRAN_ID, DIDSTD, DNAME, AMOUNT, UNIT

T2DM = type 2 diabetes mellitus.

or benefit fields within the same person-date, information was retained according to prespecified rules (Supplementary Methods S1–S4).

Laboratory Processing (Files #4)

Laboratory data were delivered in long format at the person–date–test grain (CID + DATE_SERV + LABTEST), with reported results stored in LABRESULT. LABTEST codes were mapped to analysis-friendly abbreviations and English labels using a curated lookup table. Where more than one record existed for the same CID, DATE_SERV, and mapped laboratory test, the first record in source order was retained before reshaping. The selected laboratory results were then converted from long to wide format, with one column for each mapped test. Exact duplicate wide format rows were removed. For quantitative tests, the leading numeric value was extracted from results that could contain accompanying units or text; values without an interpretable leading number and semi-quantitative strings containing terms such as “negative,” “trace,” “high,” or plus-sign categories were set to missing rather than coerced. Selected categorical results were

normalised using test-specific mappings. When multiple wide format records remained for the same CID + DATE_SERV, numeric results were averaged with zero values retained, non-missing categorical values were represented by their mode, and distinct facility or province codes were concatenated. The resulting laboratory table contained one row per CID + DATE_SERV for integration into Master_NHSO–T2DM. Quality-control checks documented record counts at each processing stage and the number of person-date records requiring collapsing.

Cross-file Merging and Demographic Harmonisation (File #1–#5)

Preprocessed visit-event tables from Files #2–#5 were merged to create a unified visit-event dataset. Because overlapping demographic and facility fields appear across multiple files, merged records were re-collapsed to ensure one row per person-date and harmonisation rules were applied. Discrepancies in sex and date of birth across sources were resolved using a consensus strategy anchored to the T2DM registry (File #1); unresolved conflicts were

set to missing to avoid potentially erroneous assignments. Harmonised demographic variables were propagated to both T2DM_registry and Master_NHSO-T2DM.

Drug Table Construction (File #6 Linked to File #5)

Drug dispensing records (File #6) were linked to service encounters (File #5) using TRAN_ID to assign CID and DATE_SERV. Drug identifiers and dose/strength were standardised using DIDSTD, the Ministry of Public Health standard drug code system [5]. Where DIDSTD was unavailable, dispensing records were harmonised using cleaned free-text drug names (Supplementary Methods S5). Drug records were then deduplicated and aggregated to a person-day-drug exposure grain defined by the composite key CID + DATE_SERV + DRUG_NAME + DOSAGE + UNIT. For records sharing this key, dispensing amount (AMOUNT) was summed, and TRAN_ID was removed from the final Drug_FINAL table.

Data Security and Governance

The dataset was de-identified by the National Health Security Office (NHSO) before transfer to the study team. Direct personal identifiers were not included, and linkage within the extract was performed using the encrypted CID and the encounter-level TRAN_ID. Data were transferred through a secure file-transfer process and are stored and analysed within secure research computing environments at the Siriraj Population Health and Nutrition Research Group, Faculty of Medicine Siriraj Hospital, Mahidol University, and the Department of Military and Community Medicine, Phramongkutklao College of Medicine, Bangkok, Thailand. Access is restricted to approved study personnel and is governed by the conditions under which the project-specific extract was provided and by the study's ethical approval from the Institutional Review Board, Royal Thai Army Medical Department (IRBRTA 1523/2024) [6]. Individual-level data are not publicly released or permitted to be shared onward without appropriate authorisation. Any proposed secondary use requires assessment of the research purpose, applicable institutional and ethics approvals, an appropriate data management and security plan, and a data use agreement specifying the permitted use and restrictions on onward sharing. No external datasets were linked in the current release. Further information on applications for controlled access is provided in the data access section.

Results

Dataset Structure and Final Analysis-Ready Outputs

The NHSO-T2DM resource is provided as a fixed ("frozen") 2017–2024 extract curated into three analysis-ready master tables (Table 2). T2DM_registry is a person-level table containing harmonised sex and birthdate, registry-recorded incident T2DM diagnosis date, and registry-defined diabetes-related diseases/outcomes grouped into four disease categories with dates. Master_NHSO-T2DM is a person-day visit-event table at the CID + DATE_SERV grain containing

merged visit-event measures, selected laboratory features, service encounter diagnoses and costs, and benefit/facility fields. Drug_FINAL is a person-day-drug exposure table containing standardised drug identifiers and dosage/strength with dispensing quantities linked to visit date; dispensing items are aligned to visit-events through linkage to service encounters. Quality-control summaries document record counts before/after processing, uniqueness at intended grains, Service-Drug linkage completeness, and missingness of key fields (Supplementary Data).

Cohort Characteristics

The extract comprises 1,421,470 individuals with registry-recorded incident T2DM during 2017–2024 in the person-level registry. Descriptive statistics reported here were generated from the analysis-ready master tables produced by the curation pipeline. The age-by-sex summary included 1,406,398 individuals with available information, of whom 812,987 (57.8%) were female. The largest age groups at recorded T2DM diagnosis were 50–59 years (467,759; 33.26%) and 60–69 years (383,880; 27.30%), together accounting for 60.56% of the age-by-sex summary population. Annual numbers of recorded incident T2DM diagnoses ranged from 163,396 in 2024 to 188,321 in 2018, demonstrating sustained cohort accrual throughout the observation period (Figure 2A and Supplementary Tables S7–S8).

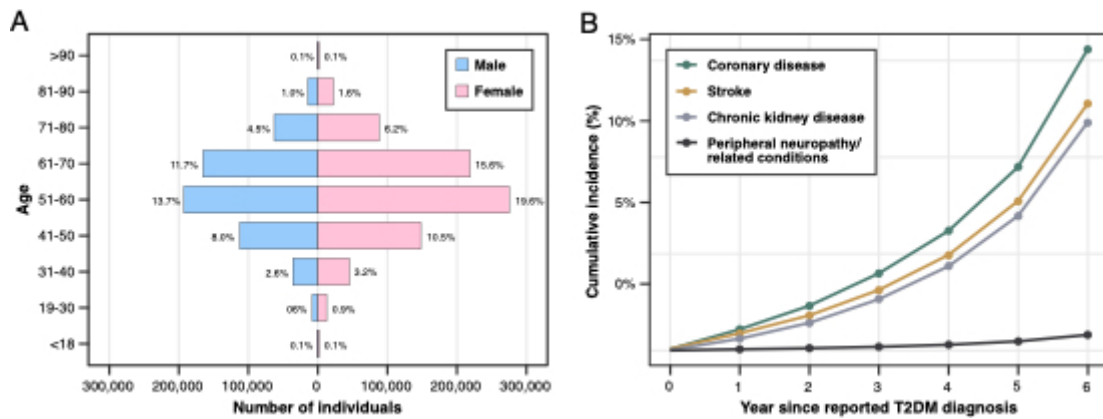
Diabetes-Related Outcomes and Interpretation of Timing

Using registry-derived dates for selected diabetes-related diseases/outcomes, cumulative incidence increases with time since recorded T2DM diagnosis (Figure 2B and Supplementary Table S10). Among the 554,471 individuals contributing five years of follow-up, the cumulative proportions with registry-recorded coronary disease, stroke, chronic kidney disease, and peripheral neuropathy/related conditions were 9.50%, 7.74%, 6.97%, and 0.47%, respectively. Estimates at longer durations are based on progressively fewer individuals with sufficient follow-up within the 2017–2024 observation window and should therefore be interpreted in relation to the shrinking population at risk over time (Supplementary Table S10). In timing summaries, diabetes-related diseases/outcomes were more frequently first recorded after the index T2DM diagnosis date, with most first records occurring ≥ 366 days post-diagnosis and smaller numbers recorded before or at diagnosis (Supplementary Table S9). Records before/at diagnosis likely reflect prevalent disease, delayed coding, or variation in recording practices, underscoring the importance of clearly defining time zero and applying baseline exclusion ("washout") windows when constructing incident outcomes from registry dates and encounter-based diagnoses.

Data Quality and Linkage Indicators

Quality-control outputs summarise within-file duplication and same-day record multiplicity prior to visit-event collapsing, and document the integrity of linkage between service encounters and dispensing items. Specifically, QC includes record counts before/after deduplication and collapsing,

Figure 2: Cohort Profile of the National Health Security Office–Type 2 Diabetes Mellitus (NHSO-T2DM) Incident Cohort (2017–2024)



Age distribution at recorded incident type 2 diabetes (T2DM) diagnosis, stratified by sex, for individuals in the analysis-ready registry table (A). Bars show counts by age band, with males shown on the left and females on the right. Cumulative incidence of registry-recorded diabetes-related diseases/outcomes by time since reported T2DM diagnosis (B). Disease categories include coronary disease, stroke, chronic kidney disease, and peripheral neuropathy/related conditions.

checks of uniqueness at intended grains for each master table, and linkage completeness metrics for Service–Drug matching via TRAN_ID (including the proportion of dispensing rows inheriting CID + DATE_SERV). Missingness summaries for key demographic fields (sex, birthdate) and selected high-value clinical/laboratory variables are provided to support appropriate interpretation and sensitivity analyses in downstream studies (Supplementary Data). All 19,413,974 distinct TRAN_ID values in the drug file matched a service record, enabling all 91,626,041 drug rows to inherit CID and DATE_SERV before deduplication and aggregation. Conversely, 5,227,561 service TRAN_ID values had no corresponding drug record, reflecting service encounters without recorded dispensing. After demographic harmonisation, sex and birthdate were available for 1,421,457 (99.9991%) and 1,405,993 (98.9112%) individuals, respectively.

Discussion

Current and Potential Uses

The NHSO–T2DM resource supports national scale longitudinal analyses of clinical trajectories, medication exposure, and diabetes-related outcomes. The combination of registry-defined cohort entry and outcome dates with repeated screening and chronic follow-up measurements, laboratory test results, encounter diagnoses and costs, and linked medication dispensing enables a range of real-world study designs, including descriptive epidemiology, pharmacoepidemiology, comparative effectiveness and safety studies, and prediction modelling [7–9]. To date, there are no peer-reviewed publications based on this specific NHSO–T2DM extract; however, the resource is currently being used in ongoing projects that illustrate its potential applications, including: (i) evaluation of visit-to-visit HbA1c variability and its association with subsequent diabetes-related complications and adverse outcomes; (ii) comparative effectiveness analyses

of commonly used second-line glucose-lowering strategies focusing on glycaemic outcomes and downstream complication risk; and (iii) development and validation of prediction models for incident diabetes complications using longitudinal routinely collected measurements, diagnoses, service utilisation, and medication dispensing data.

Strengths

This resource has several strengths for real-world T2DM research, driven by its national scale, multi-domain longitudinal capture, and consistent linkage within the extract.

- National scale and large incident cohort. The resource includes over 1.4 million individuals with registry-recorded incident T2DM during 2017–2024, enabling adequately powered analyses of relatively infrequent outcomes and robust subgroup analyses.
- Multi-domain longitudinal capture. Linkage of registry information, screening and chronic follow-up measurements, laboratory tests, encounter diagnoses and costs, and dispensed medications supports analyses of trajectories, treatment patterns, and outcome risk over time.
- Integrated, analysis-ready structure. Person-level and encounter-level linkage within the extract enables consistent integration of registry, visit-event, laboratory, diagnosis/cost, and dispensing information into unified analysis-ready tables.
- Real-world routine-care data. Data originate from routine health service delivery across diverse facilities and regions, supporting health services research and comparative effectiveness analyses in real-world practice.
- Reproducible dataset production. The transformation from delivered files to analysis-ready master tables is implemented through a documented pipeline (deduplication, visit-event collapsing, and cross-file

Table 2: Analysis-ready master tables derived from the National Health Security Office–type 2 diabetes mellitus (NHSO-T2DM) extract

Output table	Grain (unit of record)	Rows	Primary key(s)	Built from	Core contents	Typical use
T2DM_registry	Person-level	1,421,470	CID	File #1 (+ harmonised demographics from Files #2–#5)	Harmonised sex/BIRTHDATE, incident T2DM diagnosis date, diabetes-related diseases/outcomes and dates	Cohort definition; baseline characteristics; outcome anchoring
Master_NHSO-T2DM	Person-day visit-event	51,160,751	CID + DATE_SERV	File #2–#5 (+ selected labs from File #4 reshaped to wide format)	Visit-event measures, screening/exams, diagnoses, costs, facility/province/benefit fields, selected lab features	Longitudinal trajectories; feature engineering; utilisation patterns
Drug_FINAL	Person-day-drug exposure	62,229,519	CID + DATE_SERV + DRUG_NAME + DOSAGE + UNIT	File #6 linked to File #5 via TRAN_ID	Standardised drug name and dosage; dispensing amount/unit; linked to visit date	Medication exposure definitions; treatment pattern analyses

T2DM = type 2 diabetes mellitus.

demographic harmonisation), supporting consistent reuse across projects.

Limitations and Recommendations for Use

As with other routinely collected administrative and health information system data, key limitations relate to variation in completeness and coding practices, potential misclassification of exposures and outcomes, and constraints inherent to a one-time frozen extract [10, 11].

- (i) Limited clinical granularity in some domains. Some clinically important constructs are unavailable or incompletely captured, including disease severity/staging, diabetes duration prior to the first recorded T2DM date (and therefore the distinction between true incident versus newly recorded diagnosis), lifestyle factors beyond recorded fields, imaging results, and structured measures of medication adherence.
- (ii) Medication exposure is dispensing-based and requires harmonisation. Drug records reflect dispensed/claimed items rather than confirmed ingestion. Variation in coding (including free-text entries) necessitates standardisation and may leave residual misclassification, particularly for combination products and non-standard items [12]. In addition, Drug_FINAL reflects a curated medication list prioritised for T2DM care and common cardiometabolic therapies and therefore does not represent the full spectrum of dispensed medications in the source system.
- (iii) Coverage and generalisability. This extract reflects UHC beneficiaries under the NHSO and may not fully represent individuals covered under other benefit schemes; patterns of care and completeness may differ across schemes and provider networks. The dataset is also a one-time extract (2017–2024) and does not currently support routine refresh [13–15].

To support robust inference, downstream studies should consider sensitivity analyses for outcome definitions (e.g.,

registry dates and encounter diagnoses), apply baseline exclusion (“washout”) windows where appropriate when defining incident outcomes, and evaluate missingness patterns for key covariates and laboratory measures [10, 11].

Data Access

The NHSO–T2DM extract is not publicly available and cannot be shared openly by the study team. The dataset was provided by the National Health Security Office (NHSO) to the investigators as a de-identified, project-specific extract, and any access or secondary use must comply with applicable ethical approvals and the data security conditions under which the extract was provided. No digital object identifier (DOI) is currently assigned to this dataset, and it is not deposited in a public open-access repository.

Researchers interested in using or collaborating with this resource should contact the corresponding author (korapat.may@mahidol.ac.th) to discuss feasibility, scope, and the appropriate application pathway. Where access is feasible, prospective users will typically be asked to provide: (i) a brief research protocol (objectives, design, outcomes); (ii) a list of requested variables and study period; and (iii) a data management and security plan, and to obtain ethics approval from their institution (or demonstrate eligibility to work under an existing approval, as applicable). Any permitted access will be governed by an appropriate agreement specifying the permitted purpose, data protection requirements, and restrictions on onward sharing, consistent with NHSO and institutional governance requirements. Documentation (e.g., variable lists and code lists) can be provided to eligible collaborators upon request and under appropriate agreements.

Conclusions

NHSO–T2DM provides a national, linked, multi-domain routine-care data resource comprising 1,421,470 individuals with recorded incident T2DM under Thailand’s Universal Health Coverage system during 2017–2024. Its principal contribution is the transformation of six complementary

administrative and clinical source files into three documented, analysis-ready tables through explicit linkage, harmonisation, aggregation, and quality-control procedures. The resource can support reproducible studies of diabetes care patterns, medication exposure, longitudinal outcomes, comparative effectiveness and safety, and risk prediction. Findings should nevertheless be interpreted in the context of routine-data limitations, including incomplete clinical granularity, the use of dispensing records rather than confirmed medication adherence, uncertainty about whether the first recorded diagnosis represents true disease onset, and restriction to UHC beneficiaries within a fixed observation period. Controlled access through approved collaboration, ethical review, and appropriate data use agreements enables research use while protecting individual confidentiality.

Acknowledgements

We acknowledge the National Health Security Office (NHSO) for providing the data extract and thank NHSO staff for their assistance with data governance and secure transfer procedures.

Ethics Statement

This study was approved by the Institutional Review Board, Royal Thai Army Medical Department (IRBRTA 1523/2024). The dataset was de-identified prior to transfer, and informed consent was not applicable.

Conflict of Interest Statement

None declared.

Publication Consent

The NHSO–T2DM dataset is a de-identified, project-specific extract provided under ethical and data governance approvals that do not permit public release of individual-level data. We confirm that we have approval to publish this Data Resource Profile and describe the dataset and curation methods. No individual-level data are shared openly; any access is via controlled, governed processes subject to ethics approval and data use agreements.

Funding Statement

This study received no specific funding from any funding agency in the public, commercial, or not-for-profit sectors.

AI Disclosure Statement

AI-assisted tools (ChatGPT, OpenAI) were used to improve English grammar, clarity, and readability of the manuscript. All content was reviewed and edited by the authors to ensure accuracy and integrity.

Supplementary Appendices

Supplementary Appendix 1: Supplementary Data (methods, QC tables, code lists)

Data Availability Statement

The NHSO–T2DM dataset was provided by the National Health Security Office (NHSO), Thailand, as a de-identified, project-specific extract. The individual-level data are not publicly available and no DOI or public URL is assigned, due to legal, ethical, and data governance restrictions. However, the de-identified individual-level data may be shared with eligible researchers through collaboration under controlled access governance, subject to institutional approvals, ethics review, and a data use agreement that restricts onward sharing. Requests should be directed to the corresponding author, Assoc. Prof. Korapat Mayurasakorn (korapat.may@mahidol.ac.th), to discuss feasibility and the appropriate application pathway with the data custodian.

References

1. Strategy and Planning Division, Office of the Permanent Secretary, Ministry of Public Health. โครงสร้างมาตรฐานข้อมูลฯ 43 แฟ้ม [Health data standard structure (43 files)] [Internet]. Nonthaburi: Ministry of Public Health; 2025 [cited 2026 Aug 19]. Available from: <https://hdata.moph.go.th/site/>.
2. Lertpitakpong C. Strengthening health information system on hypertension: 43 folders. Public Health Policy and Laws Journal. 2021;7(3):573–582. Available from: <https://repository.li.mahidol.ac.th/entities/publication/320cf5f3-dd61-4f32-94ce-df994895a82c>.
3. National Health Security Office. ระบบส่งข้อมูลผู้ป่วยนอกและส่งเสริมป้องกันโรครายบุคคล [OP/PP individual records] [Internet]. Bangkok: National Health Security Office; 2025 [cited 2026 Aug 19]. Available from: <https://op.nhso.go.th/op/main/MainWebAction.do>.
4. Pilowsky JK, Elliott R, Roche MA. Data cleaning for clinician researchers: application and explanation of a data-quality framework. Aust Crit Care. 2024;37(5):827–833. <https://doi.org/10.1016/j.aucc.2024.03.004>
5. Strategy and Planning Division, Office of the Permanent Secretary, Ministry of Public Health. รหัสมาตรฐาน 43 แฟ้ม [Health data standards—HDC 43 files] [Internet]. Nonthaburi: Ministry of Public Health; 2025 [cited 2026 Aug 19]. Available from: https://hdata.moph.go.th/site/?page_id=84.
6. Torabi F, Squires E, Orton C, Heys S, Ford D, Lyons RA, et al. A common framework for health data governance standards. Nat Med. 2024;30(1):26–29. <https://doi.org/10.1038/s41591-023-02686-w>
7. Qu F, Shi Q, Wang Y, Shen Y, Zhou K, Pearson ER, et al. Visit-to-visit glycated hemoglobin A1c variability in

- adults with type 2 diabetes: a systematic review and meta-analysis. *Chin Med J (Engl)*. 2022;135(19):2294–2300. <https://doi.org/10.1097/CM9.0000000000002073>
8. O'Neill S, Bidulka P, Lugo-Palacios DG, Carroll O, Leiva-Escobar I, Silverwood R, et al. Comparative effectiveness of alternative second-line oral glucose-lowering therapies for type 2 diabetes: a precision medicine approach applied to routine data. *Diabetologia*. 2025;68(9):1908–1923. <https://doi.org/10.1007/s00125-025-06447-x>
 9. Ghafoor B, Zaccardi F, Khunti K, Shabnam S. Comparative predictive accuracy of machine learning versus traditional statistical methods for diabetes-related complications: a systematic review and analysis. *Diabet Med*. 2026:e70398. <https://doi.org/10.1111/dme.70398>
 10. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med*. 2015;12(10):e1001885. <https://doi.org/10.1371/journal.pmed.1001885>
 11. Emerson SD, McLinden T, Sereda P, Yonkman AM, Trigg J, Peterson S, et al. Secondary use of routinely collected administrative health data for epidemiologic research: answering research questions using data collected for a different purpose. *Int J Popul Data Sci*. 2024;9(1):2407. <https://doi.org/10.23889/ijpds.v9i1.2407>
 12. Prieto-Merino D, Mulick A, Armstrong C, Hoult H, Fawcett S, Eliasson L, et al. Estimating proportion of days covered (PDC) using real-world online medicine suppliers' datasets. *J Pharm Policy Pract*. 2021;14(1):113. <https://doi.org/10.1186/s40545-021-00385-w>
 13. Tangcharoensathien V, Patcharanarumol W, Greetong T, Suwanwela W, Kesthom N, Viriyathorn S, et al. Price setting and price regulation in health care: Thailand Universal Coverage Scheme. In: Barber SL, Lorenzoni L, Ong P, editors. Price setting and price regulation in health care: lessons for advancing universal health coverage. Geneva: World Health Organization and Organisation for Economic Co-operation and Development; 2019. p. 219–254. Available from: <https://wkc.who.int/docs/librariesprovider24/wkc-projects-documents/2-8-case-study-thailand.pdf>.
 14. Satchanawakul N, Satchanawakul N. Governance of the Universal Coverage Scheme (UCS) in Thailand [Internet]. 1st ed. Bangkok: National Health Security Office; 2023 [cited 2026 Aug 19]. Available from: https://www.nhso.go.th/images/nhsoEn/Document/UTC/Books/66-1_Governance%20Eng%20Book-%20Final16Sep.pdf.
 15. Tangcharoensathien V, editor. Thailand health system review [Internet]. 2nd ed. Manila: World Health Organization, Regional Office for the Western Pacific; 2024 [cited 2026 Aug 19]. Health Systems in Transition;13(1). ISBN: 978-92-9-062045-7. Available from: <https://apo.who.int/publications/9789290620457>.

Abbreviations

CID:	Encrypted person identifier used for deterministic linkage
CKD:	Chronic kidney disease
CSV:	Comma-separated values
DATE_SERV:	Date of service field
DIDSTD:	Thai Ministry of Public Health standard drug code
DOB:	Date of birth
DOI:	Digital object identifier
HDC:	Health Data Center
ICD-10:	International Classification of Diseases, Tenth Revision
IRBRTA:	Institutional Review Board, Royal Thai Army Medical Department
LABRESULT:	Reported laboratory result field
LABTEST:	Laboratory test code field
NCD:	Non-communicable disease
NHSO:	National Health Security Office
NHSO–T2DM:	National Health Security Office–type 2 diabetes mellitus data resource
OP/PP:	Outpatient and prevention/promotion
QC:	Quality control
T2DM:	Type 2 diabetes mellitus
TRAN_ID:	Encounter-level transaction identifier
UHC:	Universal Health Coverage

