Jurnal Kesmas Prima Indonesia
Vol. 10 No. 2 January 2026
E-ISSN : 2721-110X

Development of a Harmonized Community-Based NCD Surveillance Dataset for Public Health Research

Johannes Bastira Ginting[1], Tri Suci[2], Rizkiyah Novianti[3]

Department of Public Health, Faculty of Public Health, Universitas Prima Indonesia, Medan, Indonesia

johannesbastiraginting@unprimdn.ac.id

ABSTRACT

Community-based noncommunicable disease (NCD) surveillance data are increasingly utilized as secondary sources in public health research. Variations in dataset structure, variable nomenclature, data types, and coding systems across surveillance periods frequently impede data integration and subsequent analyses. This study developed a harmonized community-based NCD surveillance dataset using secondary data collected at UPTD Puskesmas Johar Baru, Jakarta, Indonesia, from 2019 to 2021. The methodology included dataset inventory, variable mapping, standardization of variable names, data types, and coding systems, data cleaning, derivation of analytical variables, and dataset integration using Python with the Pandas and NumPy libraries. Three surveillance datasets, comprising 196,949 observations, were successfully integrated. The harmonization process identified 158 unique variables, which were standardized into 17 analytical variables while preserving all observations. The resulting dataset offers a consistent and interoperable structure suitable for epidemiological studies, health program evaluation, and other public health research. Furthermore, the harmonization framework established in this study may serve as a practical reference for standardizing multi-period health surveillance datasets and promoting the broader use of secondary health data in future public health research.

Keywords: data harmonization; noncommunicable disease surveillance; secondary data; public health research; dataset development.

INTRODUCTION

Noncommunicable diseases (NCDs) continue to be a major public health concern because of their high morbidity, mortality, and long-term economic impact (Evyana et al., 2024; WHO, 2025). To support NCD prevention and control efforts, primary health care facilities routinely collect surveillance data on demographic characteristics, anthropometric measurements, clinical indicators, and behavioral risk factors (Riley et al., 2016; Xiong et al., 2023). These surveillance data have considerable value beyond routine reporting because they can be reused for epidemiological research, program evaluation, and evidence-informed public health policy development (Djalalinia et al., 2020).

The increasing availability of routine health surveillance data has created new opportunities for secondary data analysis and data-driven public health research (Akselrod, 2023; Djalalinia et al., 2020). However, surveillance datasets collected over multiple years often exhibit substantial heterogeneity. Differences in variable names, coding systems, data formats, measurement units, and database structures often arise because of changes in data collection practices or reporting requirements (Näher et al., 2023; Yang et al., 2022). As a result, combining datasets from different periods requires substantial preprocessing before meaningful analyses can be performed (Yang et al., 2022). Without a systematic harmonization process, inconsistencies across datasets may reduce data comparability, reproducibility, and analytical reliability (Hosseinzadeh et al., 2025; Näher et al., 2023).

Several studies have highlighted the importance of data harmonization in improving the interoperability of health information systems and facilitating the secondary use of health data (Djalalinia et al., 2020; Yang et al., 2022). Most published harmonization frameworks have been developed for large-scale health information systems, multicenter cohort studies, or international collaborative research, whereas practical harmonization procedures for routinely collected community-based surveillance data remain limited (Näher et al., 2023; Williams et al., 2023; Yang et al., 2022). To the best of our knowledge, published reports describing reproducible harmonization workflows for routinely collected community-based surveillance data at the primary health care level remain scarce, particularly within the Indonesian context. Consequently, researchers often develop study-specific data preparation procedures, resulting in inconsistent harmonization practices and limiting the reproducibility of secondary data integration across studies (Kahn et al., 2016; Yang et al., 2022).

This study aimed to develop a harmonized community-based NCD surveillance dataset using routinely collected surveillance data from UPTD Puskesmas Johar Baru, Jakarta, Indonesia, collected between 2019 and 2021. The harmonization process involved dataset inventory, variable mapping, standardization of variable names, data types, and coding systems, data cleaning, derivation of analytical variables, and integration of annual surveillance datasets into a unified analytical dataset. Beyond producing a harmonized dataset, this study proposes a systematic and reproducible harmonization workflow that can be applied to support the integration and secondary use of multi-period community-based NCD surveillance data, thereby improving the reproducibility of data harmonization for future public health research.

METHODS

Study Design

This methodological study aimed to develop a harmonized community-based noncommunicable disease (NCD) surveillance dataset by integrating routinely collected data across multiple annual surveillance periods. The study focused on developing a standardized and reproducible analytical dataset to facilitate consistent secondary data analysis while preserving the integrity of the original surveillance records.

Data Source

The study used secondary surveillance data obtained from the community-based NCD surveillance system at UPTD Puskesmas Johar Baru, Jakarta, Indonesia. The database consisted of routinely collected surveillance records from January 2019 through December 2021. Rather than selecting a sample, this study included all available surveillance records collected during the three-year surveillance period to ensure complete representation of the surveillance database and to avoid information loss during the harmonization process. Accordingly, no sampling procedure was performed.

Data Harmonization Procedure

We used a structured six-step workflow to make surveillance datasets more consistent across multiple years. First, we created a detailed inventory to find similarities and differences in variable names, definitions, coding systems, data types, and measurement units. Next, we mapped variables representing the same concept into a single data structure, ensuring their definitions matched across years. Then, we standardized variable names, data types, measurement units, and coding schemes so all datasets followed the same format. After that, we cleaned the data by fixing inconsistencies, removing duplicates or invalid records, and addressing missing or inconsistent values to improve quality. We also created new analytical variables from existing ones when needed, ensuring data were comparable across years while preserving the original information. Finally, we combined the harmonized datasets into a single standardized dataset for further public health research and verified that it was complete, consistent, and ready for use.

Evaluation of the Harmonization Process

The harmonization process was evaluated descriptively by comparing the characteristics of the original annual surveillance datasets with those of the final harmonized dataset. The harmonization workflow was evaluated using predefined process indicators, including the

number of datasets integrated, the total number of observations retained after integration, the number of unique variables identified during dataset inventory, the number of standardized analytical variables generated following harmonization, and the structural consistency of the final analytical dataset. These indicators collectively assessed the completeness, consistency, and reproducibility of the harmonization workflow rather than disease outcomes or epidemiological associations.

Ethical Considerations

This study used anonymized secondary surveillance data routinely collected as part of the community-based noncommunicable disease (NCD) surveillance program at UPTD Puskesmas Johar Baru, Jakarta, Indonesia. All personal identifiers were removed before data access, and all analyses were performed using de-identified records. The secondary use of surveillance data complied with institutional regulations governing the use of anonymized health data for research purposes.

RESULTS

Characteristics of the Original Surveillance Datasets

Three community-based noncommunicable disease (NCD) surveillance datasets collected at UPTD Puskesmas Johar Baru between 2019 and 2021 were included in the harmonization process. A total of 196,949 surveillance records were available for analysis. The structure of the annual datasets differed considerably, particularly in the number of recorded variables. The 2019 dataset contained 101 variables, whereas the 2020 and 2021 datasets contained 28 and 29 variables, respectively. These differences reflected changes in data collection and reporting practices across surveillance periods and highlighted the need for a systematic harmonization process before dataset integration.

Table 1. Characteristics of the Original Surveillance Datasets

Dataset	Year	Number of Observations	Number of Variables
Dataset 1	2019	38,936	101
Dataset 2	2020	55,372	28
Dataset 3	2021	102,641	29
Total		**196,949**	**158 unique variables***

*Unique variables identified during the dataset inventory across all surveillance years.

Variable Inventory and Harmonization

The dataset inventory identified 158 unique variable names across the three surveillance periods. Considerable variation was observed in naming conventions, coding systems, data types, and measurement formats, even though many variables represented identical concepts. A variable mapping process was subsequently performed to establish correspondence among equivalent variables and to create a standardized nomenclature. This procedure reduced inconsistencies and ensured that variables describing the same information were represented uniformly in the final dataset.

Table 2. Results of Variable Mapping and Standardization

Original Variable Name	Standardized Variable
Berat Badan	weight_kg
BB	weight_kg
Berat (Kg)	weight_kg
Tinggi Badan	height_cm
Lingkar Perut	waist_cm
Tekanan Darah Sistolik	systolic_bp
Tekanan Darah Diastolik	diastolic_bp

*****Only representative variables are presented. Complete variable mapping is available as supplementary material.

Development of the Harmonized Dataset

Following variable mapping, all datasets underwent standardization of variable names, data types, coding systems, and measurement formats before integration. Data cleaning was performed to correct inconsistencies and improve data quality, while analytical variables were derived where necessary to ensure consistency across surveillance years. The harmonization process resulted in a unified analytical dataset containing 17 standardized analytical variables without loss of surveillance records.

Table 3. Summary of the Harmonization Process

Indicator	Result
Datasets integrated	3
Total observations	196,949
Unique variables identified	158
Standardized analytical variables	17

 E-ISSN : 2721-110X

Observation retention	100%

Harmonization Workflow

The harmonization process followed a structured workflow beginning with dataset inventory and ending with the development of a unified analytical dataset. Each stage was designed to improve data consistency and ensure comparability across surveillance periods while preserving the completeness of the original surveillance records.

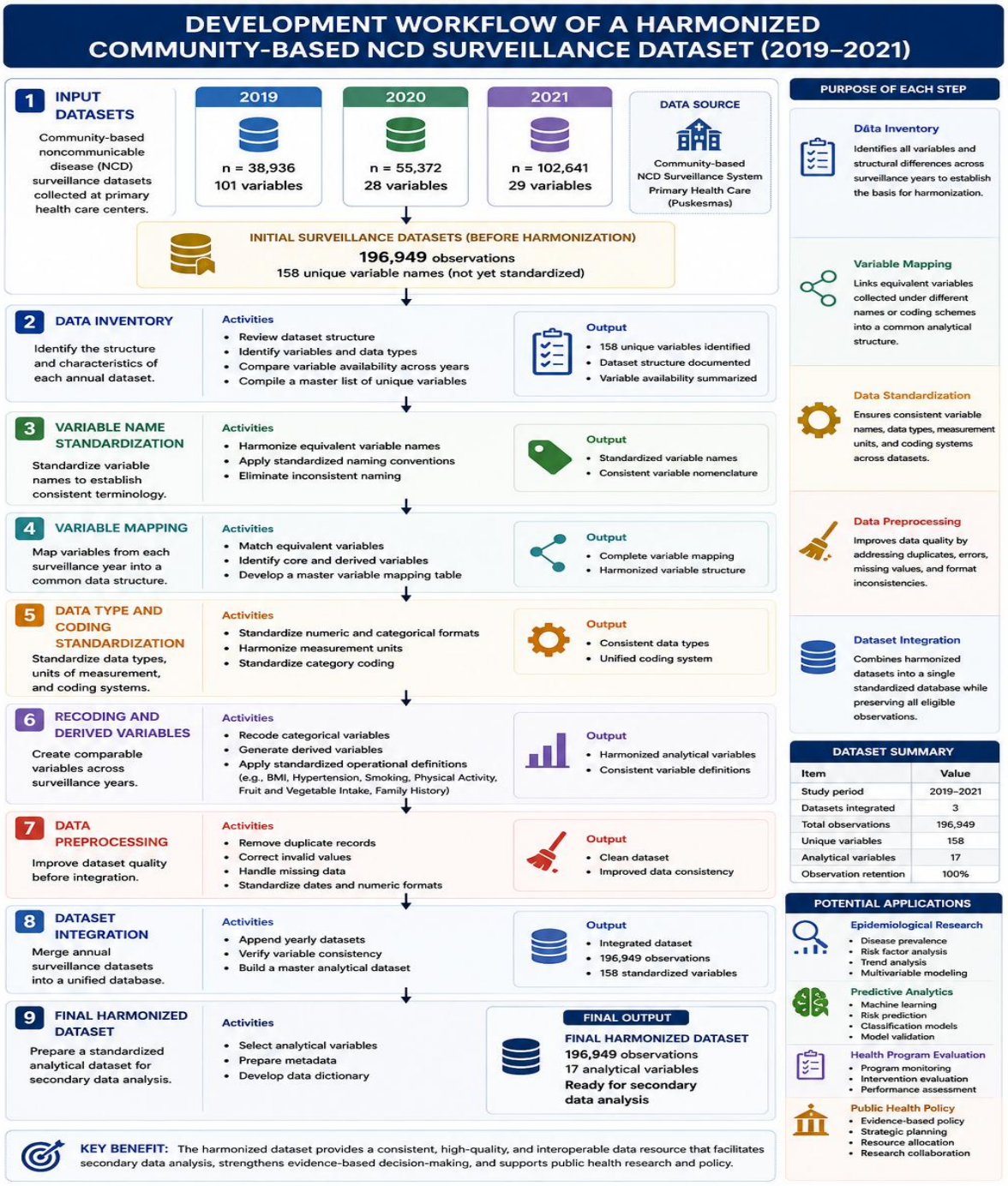

Figure 1. Workflow of the Data Harmonization Process

Characteristics of the Final Harmonized Dataset

The final harmonized dataset consisted of a standardized variable structure applicable across all surveillance years. Variable names, coding systems, data types, and measurement formats were unified, enabling direct comparison across datasets collected at different times. The resulting analytical dataset provides a consistent, interoperable data resource that can be readily used for secondary public health research without additional preprocessing.

Table 4. Comparison of Dataset Characteristics Before and After Harmonization

Component	Before Harmonization	After Harmonization
Variable names	Different across years	Standardized
Coding systems	Inconsistent	Unified
Data types	Mixed	Standardized
Dataset structure	Different	Integrated
Analytical variables	Not standardized	17 standardized variables

DISCUSSION

Principal Findings

In this study, we created a harmonized community-based noncommunicable disease (NCD) surveillance dataset by combining surveillance records collected over three years. We combined datasets in different formats into a single unified dataset, ensuring that all eligible data were retained. Our results show that using a systematic harmonization process can make surveillance data more consistent, easier to use with other systems, and more reusable for public health research. The workflow we used is practical and reproducible, so other primary health care institutions can adapt it to combine their own surveillance data from different years. We found 158 unique variables, indicating that even within the same surveillance system, data structure can change significantly from year to year. This happens because data collection methods, reporting formats, and variable definitions often change over time, which makes it harder to analyze data across multiple years. If data are not harmonized, these differences can make it difficult to compare results, repeat analyses, and prepare data for research. Our findings suggest that harmonizing data is not just a technical step, but an important part of making health

surveillance data more consistent, usable, and valuable for research (Näher et al., 2023; Yang et al., 2022).

Comparison with Previous Studies

This study's results support earlier research showing that data harmonization is key to combining diverse health datasets, improving system interoperability, and making secondary health research more reliable ((Näher et al., 2023; Yang et al., 2022). Large projects like the UK Biobank, the All of Us Research Program, and the OHDSI network have shown that using standardized data models makes it easier to share data, conduct analyses across multiple centers and groups, and reuse health data between institutions (Abbasizanjani et al., 2023; Deflaux N et al., 2023). These examples highlight how harmonized data structures help produce trustworthy evidence for clinical practice and public health decisions.

Although these large projects were built for major research environments, this study shows that the same harmonization methods can also be used for community-based surveillance data in primary health care. By following a clear six-step process—listing datasets, mapping variables, standardizing, cleaning data, creating analytical variables, and integrating datasets—different yearly surveillance datasets were combined into one unified dataset without losing valid data. This outcome suggests that systematic data harmonization is possible even in settings with limited resources, where data are mainly collected for monitoring programs rather than research. The workflow described here provides primary health care organizations with a practical, repeatable way to make their surveillance data more consistent, comparable, and useful for secondary purposes (Näher et al., 2023; Yang et al., 2022).

These findings matter beyond just creating a single dataset. The harmonized dataset and workflow provide a practical way to use community-based NCD surveillance data more effectively in public health research. Standardized datasets make it easier to track trends over time, compare data across different periods, reduce unnecessary data processing, and support a range of analyses, including epidemiological studies, health services research, predictive modeling, and machine learning. By making data more consistent and easier to combine, systematic harmonization can improve data quality and help guide public health planning and decisions. As more electronic health and surveillance data are collected, having reliable

harmonization methods like the one in this study will become even more important for making the most of these data resources(Djalalinia et al., 2020; Riley et al., 2016; Yang et al., 2022).

Methodological and Practical Implications

The development of a harmonized community-based noncommunicable disease (NCD) surveillance dataset has important methodological and practical implications for public health research and practice. A major contribution of this study is the development of a structured and reproducible data harmonization workflow specifically designed for routinely collected community-based surveillance data. Unlike many existing harmonization frameworks that primarily address national health registries, multicenter cohort studies, or large research databases, the proposed workflow demonstrates that systematic harmonization can be effectively implemented within routine primary health care surveillance systems. This extends the applicability of harmonization methodologies to resource-constrained surveillance settings where data are primarily collected for program monitoring rather than research purposes (Näher et al., 2023; Yang et al., 2022).

In addition to producing a standardized analytical dataset, this study provides transparent documentation of each harmonization stage, including dataset inventory, variable mapping, standardization, data cleaning, analytical variable derivation, and dataset integration. Explicit documentation of these procedures improves methodological transparency and reproducibility while offering a practical reference that can be adapted by researchers, public health practitioners, and health institutions seeking to standardize routinely collected surveillance data. Although the workflow was developed using surveillance data from a single primary health care facility, its underlying principles are transferable to other community-based surveillance systems with similar data structures (Djalalinia et al., 2020).

From a practical perspective, the harmonized dataset facilitates more efficient secondary use of routinely collected surveillance data by reducing data preprocessing requirements and improving data consistency across surveillance periods. Standardized datasets enable more reliable longitudinal analyses, epidemiological investigations, program monitoring and evaluation, health services research, and advanced analytical approaches such as predictive modeling and machine learning. As the volume of routinely collected electronic health and surveillance data continues to increase, practical harmonization frameworks such as the one proposed in this study can support more efficient data integration, strengthen evidence-

informed public health decision-making, and maximize the scientific value of routinely collected health data (Näher et al., 2023; Riley et al., 2016; Yang et al., 2022).

Study Limitations

Several limitations must be acknowledged when interpreting the findings of this study. First, the harmonization framework was developed from surveillance data at a single primary health care institution. The approach showed feasibility for integrating routinely collected surveillance datasets. However, other health care settings may have different database structures, variable definitions, coding systems, and reporting practices. These differences may need further adaptation before broader implementation. Therefore, generalizing the proposed framework requires caution, and additional validation with multiple datasets is necessary.

Second, this study addressed structural consistency across surveillance datasets. It did not evaluate completeness, accuracy, or validity of the original records. The quality of the harmonized dataset depends on the quality of the source data. Future research should add comprehensive data quality assessments to the harmonization process. This will enhance the reliability of secondary data analyses.

Finally, this study focused on developing a harmonized surveillance dataset. It did not assess the dataset's use in epidemiological analyses, predictive modeling, or public health program evaluation. As a result, the framework's practical utility for supporting analytical performance and decision-making remains untested. Further research is needed to evaluate how the harmonized dataset performs in public health research contexts.

CONCLUSION

A harmonized community-based noncommunicable disease (NCD) surveillance dataset was developed through a systematic process that included dataset inventory, variable mapping, standardization, data cleaning, analytical variable derivation, and dataset integration. This harmonization process enhanced the consistency, interoperability, and usability of surveillance data collected over multiple years, producing a standardized analytical dataset appropriate for secondary public health research.

In addition to developing a harmonized dataset, this study presents a practical and reproducible data harmonization workflow that can be adapted to other community-based health surveillance systems with similar data structures. The proposed framework may facilitate more efficient data integration, improve the quality and reusability of routinely collected surveillance data, and support evidence-informed public health research and decision-making. Future research

should validate the workflow using surveillance datasets from multiple regions and health information systems, and assess its applicability for epidemiological analyses, predictive modeling, and public health program evaluation.

LIMITATION

Several limitations of this study should be acknowledged. First, the proposed harmonization framework was developed using surveillance data from a single primary health care institution, which may restrict its generalizability to contexts with different database structures, coding systems, or reporting practices. Second, the harmonization process prioritized structural consistency and did not evaluate the completeness, accuracy, or validity of the original surveillance records. Consequently, the quality of the harmonized dataset remains contingent upon the underlying source data. Finally, the scope of this study was limited to the methodological development of a harmonized analytical dataset and did not include evaluation of its application in epidemiological analyses, predictive modeling, or public health program assessment. Additional validation in diverse surveillance systems is required to establish the broader applicability and practical utility of the proposed framework.

REFERENCES

Abbasizanjani, H., Torabi, F., Bedston, S., Bolton, T., Davies, G., Denaxas, S., Griffiths, R., Herbert, L., Hollings, S., Keene, S., Khunti, K., Lowthian, E., Lyons, J., Mizani, M. A., Nolan, J., Sudlow, C., Walker, V., Whiteley, W., Wood, A., … Consortium, C. C. (2023). Harmonising electronic health records for reproducible research : challenges , solutions and recommendations from a UK - wide COVID - 19 research collaboration. *BMC Medical Informatics and Decision Making*, *0*, 1–15. https://doi.org/10.1186/s12911-022-02093-0

Akselrod, S. (2023). Engaging with the Private Sector for Noncommunicable Disease Prevention and Control : Is it Possible to Create " Shared Value ?" *Annals of Global Health*, *89*(1), 1–13. https://doi.org/10.5334/aogh.4136

Deflaux N, Selvaraj, M. S., Condon, H. R., Mayo, K., Haidermota, S., Basford, M. A., Lunt, C., Philippakis, A. A., Roden, D. M., Denny, J. C., Musick, A., Collins, R., Allen, N., & Ef, M. (2023). Demonstrating paths for unlocking the value of cloud genomics through cross cohort analysis. *Nature Communications*, 1–10. https://doi.org/10.1038/s41467-023-41185-x

Djalalinia, S., Moghaddam, S. S., Sheidaei, A., & Rezaei, N. (2020). Patterns of Obesity and Overweight in the Iranian Population : Findings of STEPs 2016. *Front. Endocrinol*, *11*(February). https://doi.org/10.3389/fendo.2020.00042

Evyana, D., Novianto, E., Budianti, W. K., Krisanti, R. I. A., Wisnu, W., Wibawanti, R., Nilasari, H., Legiawati, L., Hapsari, S. A. R., & Mutmainnah, E. (2024). Association between the severity of hard-to-treat psoriasis and the prevalence of metabolic syndrome: A hospital-based cross-sectional study in Jakarta, Indonesia. *PLOS ONE*, *19*(4), e0302391. https://doi.org/10.1371/journal.pone.0302391

Hosseinzadeh, E., Afkanpour, M., & Momeni, M. (2025). Data quality assessment in healthcare , dimensions , methods and tools : a systematic review. *BMC Medical Informatics and Decision Making*.

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., & Northwest, K. P. (2016). A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data A Harmonized Data Quality Assessment Terminology and Framework for. *EDM Forum Community EGEMs*, *4*, 9–11.

Näher, A., Vorisek, C. N., Klopfenstein, S. A. I., Lehne, M., Thun, S., Alsalamah, S., Pujari, S., Heider, D., Ahrens, W., Pigeot, I., Marckmann, G., Jenny, M. A., Renard, B. Y., Kleist, M. Von, Wieler, L. H., & Balzer, F. (2023). Health Policy Secondary data for global health digitalisation. *The Lancet Digital Health*, *5*(2), e93–e101. https://doi.org/10.1016/S2589-7500(22)00195-9

Riley, L., Guthold, R., Cowan, M., Savin, S., Bhatti, L., Armstrong, T., & Bonita, R. (2016). The World Health Organization STEPwise Approach to Noncommunicable Disease Risk-Factor Surveillance : Methods , Challenges , and Opportunities. *AJPH METHODS*, *106*(1), 74–78. https://doi.org/10.2105/AJPH.2015.302962

WHO. (2025). *Noncommunicable diseases* (World Health Organization; (ed.)). World Health Organization;

Williams, E., Kienast, M., Medawar, E., Reinelt, J., Merola, A., Balzer, F., Beimes, J., B, P. Von, Chromik, J., & Arnrich, B. (2023). A Standardized Clinical Data Harmonization Pipeline for Scalable AI Application Deployment (FHIR-DHP): Validation and Usability Study Corresponding Author : *JMIR MEDICAL INFORMATICS*, *11*, 1–11. https://doi.org/10.2196/43847

Xiong, S., Cai, C., Jiang, W., Ye, P., Ma, Y., Liu, H., Li, B., Zhang, X., & Wei, T. (2023). Review Primary health care system responses to non- communicable disease prevention

and control : a scoping review of national policies in Mainland China since the 2009 health reform. *The Lancet Regional Health - Western Pacific*, *31*, 100390. https://doi.org/10.1016/j.lanwpc.2022.100390

Yang, N., Del, J., Walsh, S., Sch, C., Selby, I., Howard, K., Owen, J., Neville, J., Guiot, J., Ernst, B., Pastor, A., Alberich-bayarri, A., Menzel, M. I., Walsh, S., Vos, W., Flerin, N., Charbonnier, J., Rikxoort, E. Van, Chatterjee, A., … Uk, N. I. (2022). Data harmonisation for information fusion in digital healthcare : A state-of-the-art systematic review , meta-analysis and future research directions. *Information Fusion*, *82*(December 2021), 99–122. https://doi.org/10.1016/j.inffus.2022.01.001