pharmacoepidemiology

MDPI

Article

Pioneering Arterial Hypertension Phenotyping on Nationally Aggregated Electronic Health Records

Jing Wei Neo *, Qihuang Xie, Pei San Ang, Hui Xing Tan, Belinda Foo, Yen Ling Koon, Amelia Ng, Siew Har Tan, Desmond Teo, Mun Yee Tham, Aaron Yap, Nicholas Ng, Celine Wei Ping Loke, Li Fung Peck, Huilin Huang and Sreemanee Raaj Dorajoo

Vigilance & Compliance Branch, Health Products Regulation Group, Health Sciences Authority, Singapore 138667, Singapore
* Correspondence: neo_jing_wei@hsa.gov.sg

Abstract: Background: Hypertension is frequently studied in epidemiological studies that have been conducted using retrospective observational data, either as an outcome or a variable. However, there are few validation studies investigating the accuracy of hypertension phenotyping algorithms in aggregated electronic health record (EHR) data. Methods: Utilizing a centralized repository of inpatient EHR data from Singapore for the period of 2019–2020, a new algorithm that incorporates both diagnostic codes and medication details (Diag+Med) was devised. This algorithm was intended to supplement and improve the diagnostic code-only model (Diag-Only) for the classification of hypertension. We computed various metrics (sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV)) to assess the algorithm's effectiveness in identifying hypertension on 2813 chart-reviewed records. This pool was composed of two patient cohorts: a random sampling of all inpatient admissions (Random Cohort) and a targeted group with atrial fibrillation diagnoses (AF Cohort). Results: The Diag+Med algorithm was more sensitive at detecting hypertension patients in both cohorts compared to the Diag-Only algorithm (83.8 and 87.6% vs. 68.2 and 66.5% in the Random and AF Cohorts, respectively). These improvements in sensitivity came at minimal costs in terms of PPV reductions (88.2 and 90.3% vs. 91.4 and 94.2%, respectively). Conclusion: The combined use of diagnosis codes and specific antihypertension medication exposure patterns facilitates a more accurate capture of patients with hypertension in a database of aggregated EHRs from diverse healthcare institutions in Singapore. The results presented here allow for the bias correction of risk estimates derived from observational studies involving hypertension.

Keywords: hypertension; misclassification bias; clinical phenotyping; electronic health record; rule-based algorithm

Citation: Neo, J.W.; Xie, Q.; Ang, P.S.; Tan, H.X.; Foo, B.; Koon, Y.L.; Ng, A.; Tan, S.H.; Teo, D.; Tham, M.Y.; et al. Pioneering Arterial Hypertension Phenotyping on Nationally Aggregated Electronic Health Records. *Pharmacoepidemiology* **2024**, *3*, 169–182. https://doi.org/10.3390/pharma3010010

Academic Editor: Cristina Bosetti

Received: 19 December 2023
Revised: 1 March 2024
Accepted: 4 March 2024
Published: 12 March 2024

1. Introduction

Hypertension remains a leading risk factor for cardiovascular disease and premature death worldwide [1]. Hypertension is, thus, an important primary outcome and covariate in epidemiological studies, which are increasingly being conducted on electronic health records (EHRs). In such studies, accurately identifying patients with hypertension is a necessary first step.

However, repurposing EHR data for secondary analyses presents key challenges [2,3]. Evidence suggests tendencies towards under-coding diagnoses related to cardiovascular risk factors, such as hypertension in an individual's EHRs [4]. Using diagnosis codes alone to phenotype hypertension has been shown to result in a significant underestimation of the true disease prevalence [5].

Previous work in phenotyping hypertension has ranged from developing simple rule-based algorithms that use only hypertension-related diagnosis codes and/or antihypertensive medication exposures [6] to more complex machine-learning algorithms that

require both structured and unstructured EHR data [7]. These models have yielded acceptable sensitivity and positive predictive value (PPV) statistics on validation. However, validation has been mostly limited to data arising from the same setting as those used to develop these algorithms. The performance of any hypertension phenotyping model would expectedly vary based on prevailing setting-specific practices such as the completeness of chronic disease coding and documentation as well as the extent of capture of the prescription records for chronic medications and blood pressure measurements.

Attempting to phenotype hypertension on a nationally aggregated EHR database that draws data from different healthcare settings (from primary to tertiary care) also presents a unique challenge using different EHR systems. Serving as a consolidated repository, these aggregated databases capture individual health statuses more comprehensively. Solutions to overcome the lack of standardization upon aggregation exist, such as the conversion of EHRs to a common data model (CDM) [2], but this requires significant effort, which may not be practical. Therefore, there is still a need for the development of a broadly generalizable model that can be applied to raw aggregated EHR databases.

The primary objective of this study is to develop and validate an algorithm for predicting arterial hypertension in patients using aggregated EHR data, particularly when direct blood pressure (BP) measurements are unavailable. Such an algorithm should allow for the prevalence estimation and bias correction of risk estimates in observational studies involving hypertension. Recognizing the constraints posed by the aggregated nature of consolidated electronic health record (EHR) databases (which amalgamate data in various formats from multiple hospitals where data completeness may not be consistent), our algorithm strategically utilizes diagnostic codes and medication data to estimate hypertension status. This approach is tailored to function effectively within the limitations of the available data. Furthermore, we aim to demonstrate the feasibility of creating a robust and generalizable phenotyping algorithm that can adapt to the diverse and large datasets often encountered in EHR settings, where ideal data may not always be accessible.

2. Results

The Random Cohort was composed of 1619 inpatient admissions, with 808 patients admitted in 2019 and 811 in 2020. The mean age for this cohort was 47.5 years in 2019 and 45.8 years in 2020, reflecting a broad age distribution among the general inpatient population.

In contrast, the AF Cohort, which included patients with atrial fibrillation, consisted of 608 patients in 2019 and 586 patients in 2020. Compared to the Random Cohort, the AF Cohort had an older mean age of 72.2 years and 72.4 years, respectively, in 2019 and 2020. Additionally, there was a higher proportion of Chinese patients and a slightly greater ratio of males to females in the AF Cohort compared to the Random Cohort, as detailed in Table 1. Table A1 provides a more detailed breakdown of age by gender and race.

The two validation cohorts differed in the underlying prevalence of hypertension (Table 1). The AF Cohort had an expectedly higher prevalence (75.8 and 79.2% in 2019 and 2020, respectively) versus that of the Random Cohort (37.1 and 41.5%).

The Diag+Med hypertension algorithm was applied to both validation cohorts (Figure A1), and the results were validated via chart review. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for the two validation cohorts. The overall performance metrics of the Diag+Med hypertension algorithm were compared with the Diag-Only control algorithm in Table 2.

For both the Random and AF Cohorts, the Diag+Med algorithm outperformed the Diag-Only algorithm in sensitivity from 66.5–68.2% (Diag-Only) to 83.8–87.6% (Diag+Med). The Diag+Med algorithm also outperformed the Diag-Only algorithm in NPV, while maintaining relatively similar PPVs in both cohorts. The Diag+Med algorithm displayed lower specificity compared to the Diag-Only algorithm.

Table 1. Demographic profile of Atrial Fibrillation Cohort and Random Cohort.

		Random Cohort (n = 1619)		AF Cohort (n = 1194)	
		2019 (n = 808)	**2020 (n = 811)**	**2019 (n = 608)**	**2020 (n = 586)**
Hypertension	Yes	335 (41.5%)	301 (37.1%)	461 (75.8%)	464 (79.2%)
	No	473 (58.5%)	510 (62.9%)	147 (24.2%)	122 (20.8%)
Gender	Male	380 (47.0%)	401 (49.4%)	305 (50.2%)	310 (52.9%)
	Female	428 (53.0%)	410 (50.6%)	303 (49.8%)	276 (47.1%)
Race	Chinese	514 (63.6%)	489 (60.3%)	451 (74.2%)	458 (78.2%)
	Malay	139 (17.2%)	137 (16.8%)	92 (15.1%)	81 (13.8%)
	Indian	84 (10.4%)	99 (12.3%)	29 (4.8%)	25 (4.3%)
	Others	71 (8.8%)	86 (10.6%)	36 (5.9%)	22 (3.8%)
Age	Mean	47.5	45.8	72.2	72.4
	Standard deviation	28.8	27.5	11.8	12.0
Total		808 (100.0%)	811 (100.0%)	608 (100.0%)	586 (100.0%)

AF: atrial fibrillation.

Table 2. Comparison of Diag+Med algorithm performance with Diag-Only algorithm performance.

	Diag-Only (%)		Diag+Med (%)	
Cohort	**Random Cohort (n = 1619)**	**AF Cohort (n = 1194)**	**Random Cohort (n = 1619)**	**AF Cohort (n = 1194)**
Sensitivity	68.2	66.5	83.8	87.6
Specificity	95.8	85.9	92.8	67.7
PPV	91.4	94.2	88.2	90.3
NPV	82.3	42.7	89.9	61.3

Diag-Only: diagnosis code-only algorithm (from diagnosis and Patient Problem List). Diag+Med: diagnosis code and medication algorithm (Figure 1). AF: atrial fibrillation; PPV: positive predictive value; NPV: negative predictive value.

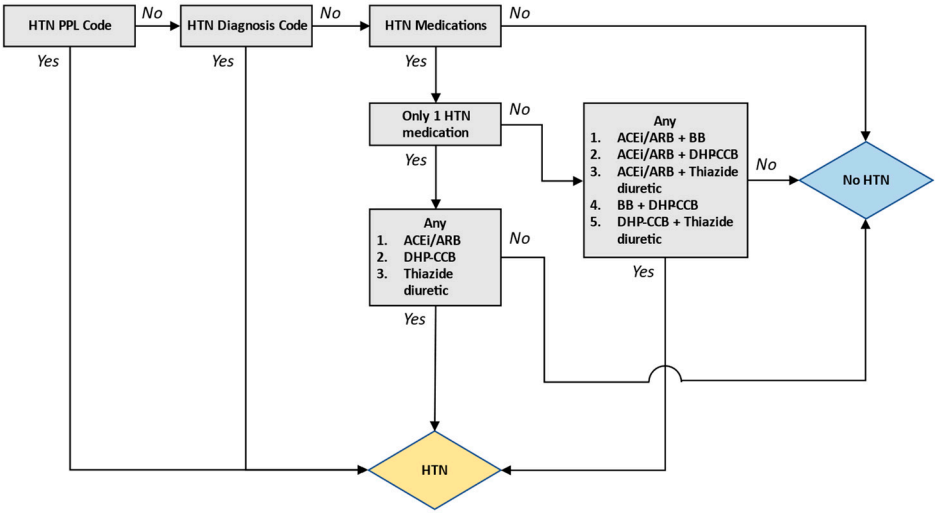

Figure 1. Diagnosis- and medication-based hypertension phenotyping algorithm flow chart (Diag+Med algorithm). PPL: Patient Problem List; ACEi: angiotensin-converting enzyme inhibitor; ARB: angiotensin receptor blocker; BB: beta blocker; DHP-CCB: dihydropyridine-calcium channel blocker.

The year-wise performance metrics by cohort (2019 and 2020) of the Diag+Med algorithm are shown in Table 3. In the Random Cohort, the Diag+Med algorithm had a sensitivity of 82.45–85.4% (2019 and 2020, respectively), specificity of 92.0–93.5%, PPV of 87.95–88.6%, and NPV of 88.1–91.6%. In the AF Cohort, the algorithm had a sensitivity of 85.9–89.2%, specificity of 66.7–68.9%, PPV of 89.0–91.6%, and NPV of 60.1–62.7%. Performance metrics were also calculated across years of admission, gender, and race (Table A2).

Table 3. Performance of the Diag+Med algorithm on the AF Cohort and the Random Cohort.

	Random Cohort (n = 1619)			AF Cohort (n = 1194)		
Statistics	**2019 (%)**	**2020 (%)**	**Overall (%)**	**2019 (%)**	**2020 (%)**	**Overall (%)**
Sensitivity	82.4	85.4	83.8	85.9	89.2	87.6
Specificity	92.0	93.5	92.8	66.7	68.9	67.7
PPV	87.9	88.6	88.2	89.0	91.6	90.3
NPV	88.1	91.6	89.9	60.1	62.7	61.3

AF: atrial fibrillation; PPV: positive predictive value; NPV: negative predictive value.

3. Discussion

This retrospective validation study, conducted on Singaporean hospital inpatients, demonstrated an overall good performance in phenotyping patients with arterial hypertension. The performance of our Diag+Med algorithm was comparable to other hypertension phenotype algorithms developed in other countries. These other studies similarly highlighted that diagnosis codes were usually only inadequate for capturing hypertension [5,7]. An example is an American study by Teixeira et al. in which they developed multiple algorithms using a combination of diagnosis codes, medications, and BP measurements, achieving sensitivity and PPV of above 80–90% [7].

However, the crux of our Diag+Med algorithm lies in its useability on heterogeneous aggregated EHR databases without requiring BP measurements, which is the primary objective of our study. Teixeira's phenotyping algorithms were developed using data from only one hospital cluster [7], in contrast to our Diag+Med algorithm, which was developed and tested on aggregated data from multiple hospital clusters in Singapore and did not require BP measurements.

Another key advantage of our Diag+Med algorithm is that it operates directly on raw EHR data without requiring the laborious process of conversion to a CDM, unlike other rule-based hypertension phenotyping algorithms [8].

This algorithm is a pioneering study in the use of Singapore's nationwide EHRs to phenotype patients. The strength of this study lies in the novelty and scale of the EHR data accessed (covering approximately 85% of all hospital admissions in Singapore [9]), as well as the large and varied cohorts used to evaluate the generalizability of the proposed algorithm. The algorithm validation involved a relatively large sample size of different patient profiles from multiple contributing healthcare institutions, showing that the algorithm is fit for use on diverse patient populations from different healthcare clusters. This facilitates the identification of patients with hypertension from aggregated data sources in Singapore without the need for additional harmonization or processing (such as conversion to a CDM).

Hypertension is a chronic condition that is usually managed on an outpatient basis [10]; therefore, physicians may not input hypertension as a diagnosis for an inpatient admission. This illustrates the importance of including patient medication data in the phenotyping of hypertension, as it considerably improves sensitivity without excessively sacrificing specificity.

Compared to the AF Cohort, the Diag+Med algorithm was better at distinguishing negative cases in the Random Cohort. This was likely due to the difference in patient profiles between the two cohorts, with the AF Cohort having a much higher underlying

prevalence of hypertension. Higher hypertension prevalence in the AF Cohort resulted in lower specificity [11]. Evidence also suggests that increased prevalence may result in a variance in sensitivity and specificity even though these measures are theoretically independent of prevalence, possibly due to other mechanisms [12].

The Diag+Med algorithm performed consistently over consecutive years (2019 and 2020, shown in Table 3), with all performance metrics (sensitivity, specificity, PPV, and NPV) varying less than 5% across the years for all validation cohorts. The data demonstrate the algorithm's stability throughout the 2020 period, indicating resilience to any potential impact caused by the COVID-19 pandemic. This is critical as it suggests that the algorithm can reliably function under varying conditions, a feature that is essential for real-world applications.

A sub-group analysis of the Diag+Med algorithm across gender and race was also conducted (Table A2). The algorithm performed consistently across genders and most races in Singapore. However, caution should be taken in interpreting the results of the sub-group analysis. The findings may be attributable to chance, especially since the validation study was not specifically designed to focus on these sub-group analyses, and certain sub-groups are relatively small (such as the Others ethnicity as listed in Table 1).

There are some limitations to the EHR database available to researchers. Due to the nature of the EHR database, which lacked vital sign readings such as BP readings, it was crucial to develop an algorithm that was able to phenotype hypertension without such data. There remains a need for an algorithm that can estimate the prevalence of hypertension and adjust risk estimates in epidemiological studies using the available data. Our algorithm is designed to work within these constraints, leveraging diagnostic codes and medication data to provide the best possible estimation of hypertension status in the absence of direct BP measurements. It is notable that in this study, BP readings were not needed to produce a hypertension algorithm with a good performance.

Our study was developed and validated on hospital inpatients. Due to the nature of our database, which contained unstructured notes from inpatient settings but not outpatient settings, our ability to carry out comprehensive chart reviews on non-hospitalized patients was limited. Inpatient cases would have their past medical history extensively documented in the discharge summary, but not outpatients; hence, it was not possible to carry out a robust algorithm validation on an outpatient study cohort.

Furthermore, the algorithm may not accurately predict patients who are followed up in private settings, such as by general practitioners (GPs) or in private hospitals, as their medications and outpatient visits are not available in the database. This likely contributed to the false negative cases in the algorithm's validation as there is no visibility of the patient data outside of their inpatient admissions.

In Singapore, some medications for hypertension are commonly prescribed for other conditions, such as heart failure and coronary heart syndrome (e.g., angiotensin II receptor blockers and ACE inhibitors) [13]. This potentially contributes to the false positive rate in the algorithm, and further improvements to the algorithm should consider the presence of such comorbidities in addition to patients' medication lists.

The performance of the algorithm may vary over time as hypertension prescribing guidelines, coding practices, and EHR systems in public hospitals may change in the future. Caution must be taken to ensure that these underlying trends are stable before applying the hypertension algorithm to cohorts.

4. Materials and Methods

4.1. Data Sources

All available historical records were extracted from a database that contains aggregated, de-identified clinical data from all public healthcare institutions in Singapore. This database covers approximately 85% of all hospital admissions and over 40% of all chronic outpatient visits [8]. The database did not undergo prior harmonization or processing (e.g., conversion to a CDM). Structured clinical data include patient demographics, diagnosis

codes in SNOMED (Systematized Nomenclature of Medicine), ICD-10 (International Classification of Diseases, 10th Revision) formats, and dispensed medication records from both outpatient and inpatient settings. Unstructured clinical data include hospital discharge summaries and emergency department visit notes.

Diagnosis codes were extracted from two data tables: (1) diagnosis and (2) Patient Problem List (PPL). A Patient Problem List includes active issues with current management, background chronic conditions, and resolved past medical issues. Medications dispensed at the outpatient or inpatient discharge pharmacies from all contributing data centers were extracted from the medications table.

Data elements from structured and unstructured clinical data (such as discharge summaries and laboratory tests) were available for a chart review; however, vital sign readings such as blood pressure (BP) were not accessible.

4.2. Algorithm Development and Validation

The primary outcome of interest was defined as chronic arterial hypertension, with the exclusion of pulmonary hypertension, pre-eclampsia/gestational hypertension, ocular hypertension, peripheral venous hypertension, and portal hypertension. Diagnosis codes from the diagnosis and PPL tables, and medications from the medications table, were used to develop a combined diagnosis and medication data-based hypertension phenotyping algorithm, as shown in Figure 1 (Diag+Med algorithm). A diagnosis code-only algorithm, which exclusively relied on the presence of diagnosis codes, was used as a control (Diag-Only algorithm), as shown in Figure 2.

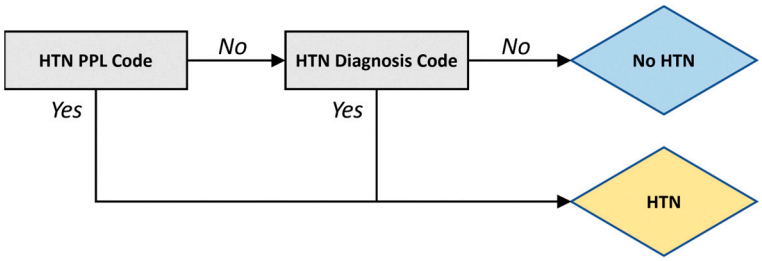

Figure 2. Diagnosis code-based hypertension phenotyping algorithm flow chart (Diag-Only algorithm). PPL: Patient Problem List.

A broad list of candidate SNOMED and ICD diagnosis codes for hypertension was first identified from ICD-9 AM, ICD-10, and SNOMED CT browsers. A frequency of use assessment was conducted to identify commonly used codes; diagnosis codes with fewer than 10 patients found in the database between 2018 and 2021 were removed. A hospital physician vetted the remaining diagnosis codes and descriptions to ensure their appropriateness for identifying chronic hypertension (Table A3). First- and second-line medications used to manage hypertension were shortlisted based on the American Heart Association's 2017 [14] and the Ministry of Health (Singapore)'s 2017 guidelines [9] (Table A4). Patients treated with beta blockers alone were not included. The diagnosis codes and medications listed in Tables A3 and A4 were used in the phenotyping algorithms.

For the Diag+Med algorithm (Figure 1), patients were categorized as hypertensive if they had any PPL or diagnosis table records with a diagnosis code found in Table A3. If the patient did not have any diagnosis codes in Table A3, the algorithm would look at the medications prescribed to the patient. Patients were classified as hypertensive if they were prescribed one medication listed in Table A4, specifically an angiotensin-converting enzyme inhibitor (ACEi), angiotensin receptor blocker (ARB), dihydropyridine-calcium channel blocker (DHP-CCB), or thiazide diuretic. Alternatively, patients were also classified as hypertensive if they were prescribed a combination of any two of the following medications from Table A4: (a) ACEi or ARB with a beta blocker (BB), (b) ACEi or ARB with a DHP-CCB,

(c) ACEi or ARB with a thiazide diuretic, (d) BB with a DHP-CCB, or (e) DHP-CCB with a thiazide diuretic.

For the Diag-Only algorithm (Figure 2), patients were categorized as hypertensive if they had any PPL or diagnosis table records with a diagnosis code found in Table A3.

The hypertension algorithms (Diag+Med and Diag-Only) were applied on two validation cohorts (Random Cohort and AF Cohort). These validation cohorts were constructed by sampling patients admitted to any public health institution between 2019 and 2020, as shown in Figure 3. The Random Cohort consists of a random sample of inpatient admissions in 2019 or 2020 from any public health institution. Random sampling of the dataset was necessitated by the constraints of our available computational resources. The AF Cohort consists of inpatient admissions with a new diagnosis of atrial fibrillation (AF) in 2019 and 2020. The inclusion criteria for the AF Cohort were defined as a new onset of primary or secondary diagnoses of AF (an ICD-10 or SNOMED diagnosis code of atrial fibrillation) and the initiation of one of the drugs of interest (apixaban, rivaroxaban, or warfarin) within 2 days before the date of discharge.

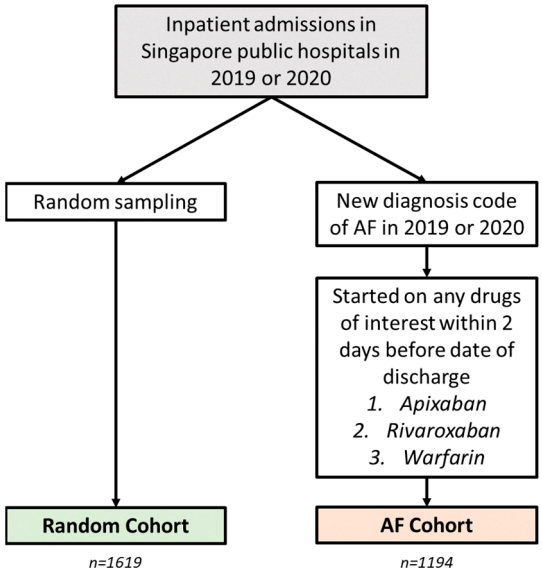

Figure 3. Construction of Random Cohort and Atrial Fibrillation (AF) Cohort.

The Random Cohort was used to assess the algorithm's generalizability and performance in a diverse patient population. Additionally, the AF Cohort was chosen due to the higher prevalence of hypertension within this group compared to the general inpatient population, providing a robust test for the algorithm's sensitivity in a group with higher prevalence.

Trained annotators from the Health Sciences Authority independently assessed all sampled admissions from both validation cohorts via a chart review of the aggregated database. Only data that were recorded before or on the discharge date of the patient's inpatient admission episode were used.

A trial annotation run-in phase was conducted for annotators to practice annotating for hypertension on a common set of 200 patient charts (not included in this study) to assess potential variability in annotation accuracy. An excellent inter-annotator agreement of 0.89–0.99 was achieved on the 200 practice set records (pairwise Cohen's Kappa, Table A5). Thereafter, independent (non-overlapping) annotations were carried out on all records in both AF and Random validation cohorts to develop a gold-standard label for each patient in the validation cohorts.

4.3. Statistical Analysis

Algorithm performance metrics (sensitivity, specificity, PPV, and NPV) were calculated by comparing the hypertension predictions of the algorithm with the gold-standard labels reviewed in the patient charts during annotation. Sensitivity was calculated by taking the proportion of confirmed hypertension cases that were predicted to be positive (true positives) out of all confirmed hypertension cases (true positives and false negatives). Specificity was calculated by taking the proportion of confirmed non-hypertension cases that were predicted to be negative (true negatives) out of all confirmed non-hypertension cases (true negatives and false positives). PPV was calculated by taking the proportion of confirmed hypertension cases that were predicted to be positive (true positives) out of all cases that were predicted positive (true positives and false positives). NPV was calculated by taking the proportion of confirmed non-hypertension cases that were predicted to be negative (true negatives) out of all cases that were predicted negative (true negatives and false negatives). The algorithm's performance metrics and Cohen's Kappa were calculated using Spyder (Python 3.8).

5. Conclusions

The development and validation of a hypertension phenotyping algorithm with high sensitivity and specificity were not only beneficial for identifying hypertensive patients in various clinical and pharmacoepidemiology studies, but they also demonstrated its effectiveness in the context of Singapore. This algorithm is particularly noteworthy as it can be successfully applied to national aggregated data sourced from diverse healthcare institutions across the country, without requiring harmonization or conversion to a CDM. This makes it a versatile and robust tool for the identification of patients with hypertension within the healthcare landscape in Singapore to facilitate risk estimate adjustments or quantitative bias analysis in epidemiological studies.

Author Contributions: J.W.N., Q.X. and S.R.D. designed the study. J.W.N. and Q.X. analyzed the data. J.W.N., P.S.A., H.X.T., B.F., Y.L.K., A.N., S.H.T., D.T., M.Y.T., A.Y., N.N., C.W.P.L., L.F.P., H.H. and S.R.D. provided the domain expertise for the manual annotation of discharge summaries. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The information used for this study is not available in the public domain. The code is not available as it has been written for specific fields in the dataset. Interested parties may refer to the algorithm's logic, which is included in this paper.

Acknowledgments: Yan Tong Loo vetted diagnosis codes for use in the hypertension phenotyping algorithm.

Conflicts of Interest: The authors have no conflicts of interest that are directly relevant to the content of this article. The views expressed in this article may not be understood or quoted as being made on behalf of or reflecting the position of HSA.

Appendix A

Table A1. Demographic Profile of Atrial Fibrillation Cohort and Random Cohort, with breakdown by gender and race.

		Random Cohort (n = 1619)		AF Cohort (n = 1194)	
		2019 (n = 808)	**2020 (n = 811)**	**2019 (n = 608)**	**2020 (n = 586)**
Hypertension	Yes	335 (41.5%)	301 (37.1%)	461 (75.8%)	464 (79.2%)
	No	473 (58.5%)	510 (62.9%)	147 (24.2%)	122 (20.8%)
Gender	Male	380 (47.0%)	401 (49.4%)	305 (50.2%)	310 (52.9%)
	Female	428 (53.0%)	410 (50.6%)	303 (49.8%)	276 (47.1%)
Race	Chinese	514 (63.6%)	489 (60.3%)	451 (74.2%)	458 (78.2%)
	Malay	139 (17.2%)	137 (16.8%)	92 (15.1%)	81 (13.8%)
	Indian	84 (10.4%)	99 (12.3%)	29 (4.8%)	25 (4.3%)
	Others	71 (8.8%)	86 (10.6%)	36 (5.9%)	22 (3.8%)
Age (Mean, SD)	Overall	47.5 (28.8)	45.8 (27.5)	72.2 (11.8)	72.4 (12.0)
	Male	47.5 (29.9)	47.0 (28.4)	69.1 (11.6)	69.7 (12.0)
	Female	47.5 (27.9)	44.6 (26.4)	75.3 (11.2)	75.3 (11.4)
	Chinese	52.8 (28.1)	52.9 (27.8)	73.8 (11.0)	73.7 (10.9)
	Malay	33.2 (26.7)	31.6 (24.8)	67.6 (15.6)	69.6 (13.7)
	Indian	45.8 (27.4)	39.3 (22.9)	69.0 (11.0)	66.7 (13.6)
	Others	39.1 (28.7)	35.2 (19.5)	63.9 (14.6)	67.8 (18.3)
Total		808 (100.0%)	811 (100.0%)	608 (100.0%)	586 (100.0%)

AF: atrial fibrillation.

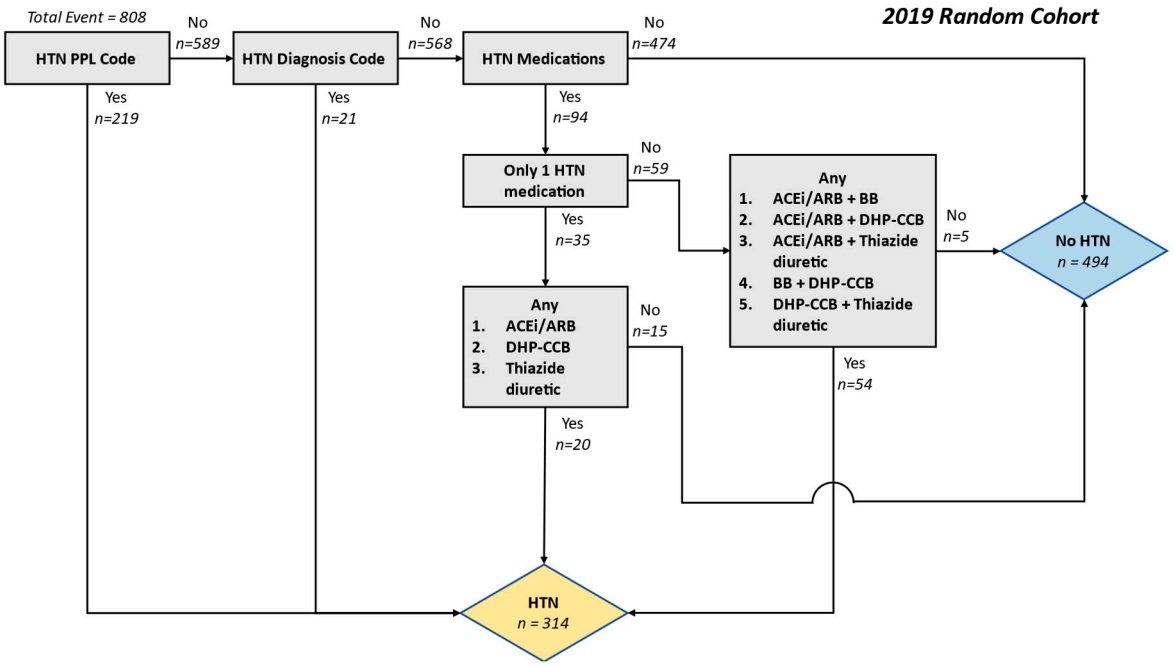

Figure A1. Illustrative flow chart of Diag+Med algorithm applied to 2019 Random Cohort. PPL: Patient Problem List; ACEi: angiotensin-converting enzyme inhibitor; ARB: angiotensin receptor blocker; BB: beta blocker; DHP-CCB: dihydropyridine-calcium channel blocker.

Table A2. Breakdown of Diag+Med algorithm performance compared with Diag-Only algorithm performance by year of admission, gender, and race.

		Diag-Only (%)		Diag+Med (%)	
		Random Cohort (n = 1619)	**AF Cohort (n = 1194)**	**Random Cohort (n = 1619)**	**AF Cohort (n = 1194)**
Sensitivity	Overall	68.2	66.5	83.8	87.6
	2019	65.1	63.1	82.4	85.9
	2020	71.8	69.8	85.4	89.2
	Male	67.1	68.3	82.4	87.7
	Female	69.7	64.6	85.5	87.4
	Chinese	68.7	68.1	84.3	88.6
	Malay	71.2	58.4	86.4	84.7
	Indian	76.5	68.6	88.2	88.6
	Others	45.2	63.4	66.7	78.0
Specificity	Overall	95.8	85.9	92.8	67.7
	2019	95.3	85.7	92.0	66.7
	2020	96.3	86.1	93.5	68.9
	Male	95.0	84.1	91.9	65.6
	Female	96.5	88.1	93.5	70.3
	Chinese	93.9	85.3	90.2	67.0
	Malay	98.1	86.1	95.7	75.0
	Indian	98.3	84.2	95.7	57.9
	Others	98.3	94.1	96.5	70.6
PPV	Overall	91.4	94.2	88.2	90.3
	2019	90.8	93.3	87.9	89.0
	2020	91.9	95.0	88.6	91.6
	Male	91.3	93.0	88.8	88.7
	Female	91.4	95.5	87.6	92.0
	Chinese	90.5	94.4	88.0	90.7
	Malay	92.2	94.1	86.4	92.8
	Indian	96.3	88.9	92.3	79.5
	Others	90.5	96.3	87.5	86.5
NPV	Overall	82.3	42.7	89.9	61.3
	2019	79.4	42.6	88.1	60.1
	2020	85.2	42.9	91.6	62.7
	Male	78.7	46.4	87.0	63.5
	Female	85.5	39.0	92.3	58.9
	Chinese	78.0	42.5	87.2	62.0
	Malay	91.6	35.2	95.7	56.2
	Indian	87.6	59.3	93.2	73.3
	Others	83.1	51.6	88.8	57.1

AF: atrial fibrillation.

Table A3. Diagnosis codes used in hypertension phenotyping algorithm.

No.	Diagnosis Code	Diagnosis Description	Format
1	38341003	Hypertensive disorder	SNOMED
2	59621000	Essential hypertension	SNOMED
3	10725009	Benign hypertension	SNOMED
4	38481006	Hypertensive renal disease	SNOMED
5	1201005	Benign essential hypertension	SNOMED
6	6962006	Hypertensive retinopathy	SNOMED
7	64715009	Hypertensive heart disease	SNOMED
8	56218007	Systolic hypertension	SNOMED
9	170578008	Poor hypertension control	SNOMED
10	I10	Essential (primary) hypertension	ICD-10
11	86041002	Pre-existing hypertension in obstetric context	SNOMED
12	86234004	Hypertensive heart AND renal disease	SNOMED
13	473392002	Hypertensive nephrosclerosis	SNOMED
14	266287006	(Hypertensive disease) or (hypertension)	SNOMED
15	8762007	Chronic hypertension in obstetric context	SNOMED
16	712832005	Supine hypertension	SNOMED
17	5148006	Hypertensive heart disease with congestive heart failure	SNOMED
18	65402008	Pre-existing hypertension complicating AND/OR reason for care during pregnancy	SNOMED
19	78975002	Malignant essential hypertension	SNOMED
20	194779001	Hypertensive heart and renal disease with (congestive) heart failure	SNOMED
21	46113002	Hypertensive heart failure	SNOMED
22	48146000	Diastolic hypertension	SNOMED
23	194767001	Benign hypertensive heart disease with congestive cardiac failure	SNOMED
24	397748008	Hypertension with albuminuria	SNOMED
25	49220004	Hypertensive renal failure	SNOMED
26	443482000	Hypertensive urgency	SNOMED
27	62275004	Hypertensive episode	SNOMED
28	50490005	Hypertensive encephalopathy	SNOMED
29	706882009	Hypertensive crisis	SNOMED
30	70272006	Malignant hypertension	SNOMED
31	31992008	Secondary hypertension	SNOMED
32	161501007	H/O: hypertension *	SNOMED
33	52698002	Transient hypertension	SNOMED
34	123799005	Renovascular hypertension	SNOMED
35	28119000	Renal hypertension	SNOMED
36	193003	Benign hypertensive renal disease (disorder)	SNOMED
37	194785008	Benign secondary hypertension	SNOMED
38	449759005	Hypertensive complication	SNOMED
39	428163005	Hypertensive left ventricular hypertrophy	SNOMED
40	89242004	Malignant secondary hypertension	SNOMED
41	37618003	Chronic hypertension complicating AND/OR reason for care during pregnancy	SNOMED

* History of hypertension.

Table A4. Medications used in hypertension phenotyping algorithm.

No.	Class of Medicine Included	ATC L4 Code	Included Drugs (Not Exclusive)	Excluded Drugs
1	Dihydropyridine derivatives	C08CA C08GA	Amlodipine Nifedipine Felodipine Lacidipine Cilnidipine Nimodipine	
2	Angiotensin II antagonists, plain	C09CA	Losartan Valsartan Telmisartan Irbesartan Candesartan Olmesartan medoxomil	
3	ACE inhibitors, plain	C09AA	Enalapril Lisinopril Perindopril Captopril Ramipril Imidapril	
4	Beta blocking agents, selective	C07AB	Atenolol Bisoprolol Metoprolol Nebivolol	Sotalol Timolol Betaxolol Esmolol
5	Angiotensin II antagonists and calcium channel blockers	C09DB C09DX	Valsartan and amlodipine Telmisartan and amlodipine Olmesartan, medoxomil, and amlodipine	
6	Thiazides, plain	C03AA	Hydrochlorothiazide	
7	Sulfonamides, plain	C03BA C03CA	Furosemide Indapamide Metolazone Bumetanide	Verapamil
8	Alpha and beta blocking agents	C07AG	Carvedilol Labetalol	
9	Organic nitrates	C01DA C01DB C01DX	Isosorbide dinitrate Isosorbide mononitrate	Glyceryl trinitrate
10	Beta blocking agents, non-selective	C07AA	Propranolol Nadolol	
11	Angiotensin II antagonists and diuretics	C09DA	Valsartan and diuretics Losartan and diuretics Irbesartan and diuretics	
12	Benzothiazepine derivatives	C08DB	Diltiazem	
13	Aldosterone antagonists	C03DA	Spironolactone Eplerenone	
14	Beta blocking agents, selective, and other antihypertensives	C07FB C07FX	Atenolol and other antihypertensives	
15	ACE inhibitors and calcium channel blockers	C09BB	Perindopril and amlodipine	
16	Low-ceiling diuretics and potassium-sparing agents	C03EA	Hydrochlorothiazide and potassium-sparing agents	
17	ACE inhibitors, other combinations	C09BX	Perindopril, amlodipine and indapamide Cosyrel	
18	Angiotensin II antagonists, other combinations	C09DX	Sacubitril-valsartan	
NA	Other excluded medicines	C03XA C01DX		Tolvaptan Nicorandil

Table A5. Inter-annotator agreement among annotators (Cohen's Kappa). Pairwise Cohen's Kappa score between fifteen annotators on training dataset of 200 discharge summaries.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1		0.99	0.99	0.96	0.97	0.95	0.97	0.97	0.92	0.96	0.98	0.93	0.96	0.99	0.96
2	0.99		0.98	0.95	0.96	0.96	0.98	0.96	0.93	0.95	0.99	0.94	0.95	0.98	0.97
3	0.99	0.98		0.97	0.98	0.94	0.96	0.98	0.93	0.95	0.97	0.94	0.97	0.98	0.95
4	0.96	0.95	0.97		0.95	0.91	0.93	0.95	0.9	0.92	0.96	0.95	0.96	0.95	0.92
5	0.97	0.96	0.98	0.95		0.92	0.94	0.96	0.91	0.95	0.95	0.92	0.95	0.96	0.93
6	0.95	0.96	0.94	0.91	0.92		0.94	0.92	0.91	0.91	0.95	0.92	0.95	0.96	0.93
7	0.97	0.98	0.96	0.93	0.94	0.94		0.94	0.91	0.93	0.97	0.92	0.93	0.96	0.95
8	0.97	0.96	0.98	0.95	0.96	0.92	0.94		0.91	0.93	0.95	0.92	0.95	0.96	0.93
9	0.92	0.93	0.93	0.90	0.91	0.91	0.91	0.91		0.90	0.92	0.89	0.92	0.91	0.90
10	0.96	0.95	0.95	0.92	0.95	0.91	0.93	0.93	0.90		0.94	0.89	0.92	0.95	0.92
11	0.98	0.99	0.97	0.96	0.95	0.95	0.97	0.95	0.92	0.94		0.95	0.94	0.97	0.96
12	0.93	0.94	0.94	0.95	0.92	0.92	0.92	0.92	0.89	0.89	0.95		0.91	0.94	0.91
13	0.96	0.95	0.97	0.96	0.95	0.95	0.93	0.95	0.92	0.92	0.94	0.91		0.95	0.92
14	0.99	0.98	0.98	0.95	0.96	0.96	0.96	0.96	0.91	0.95	0.97	0.94	0.95		0.95
15	0.96	0.97	0.95	0.92	0.93	0.93	0.95	0.93	0.90	0.92	0.96	0.91	0.92	0.95	

References

1. Brouwers, S.; Sudano, I.; Kokubo, Y.; Sulaica, E.M. Arterial hypertension. *Lancet* **2021**, *10296*, 249–261. [CrossRef] [PubMed]
2. Ta, C.N.; Weng, C. Detecting Systemic Data Quality Issues in Electronic Health Records. *Stud. Health Technol. Inform.* **2019**, *264*, 383–387. [CrossRef] [PubMed]
3. D'Amore, J. Electronic Health Record Data Governance and Data Quality in the Real World. Healthcare Information and Management Systems Society. 2023. Available online: https://www.himss.org/resources/electronic-health-record-data-governance-and-data-quality-real-world (accessed on 16 February 2023).
4. Angelow, A.; Reber, K.C.; Schmidt, C.O.; Baumeister, S.E.; Chenot, J.-F. Prevalence of Cardiovascular Risk Factors at The Population Level: A Comparison of Ambulatory Physician-Coded Claims Data with Clinical Data from A Population-Based Study. *Gesundheitswesen* **2019**, *81*, 791–800. [CrossRef] [PubMed]
5. Peng, M.; Chen, G.; Kaplan, G.G.; Lix, L.M.; Drummond, N.; Lucyk, K.; Garies, S.; Lowerison, M.; Weibe, S.; Quan, H. Methods of defining hypertension in electronic medical records: Validation against national survey data. *J. Public Health* **2016**, *38*, e392–e399. [CrossRef] [PubMed]
6. Nadkarni, G.N.; Gottesman, O.; Linneman, J.G.; Chase, H.; Berg, R.L.; Farouk, S.; Nadukuru, R.; Lotay, V.; Ellis, S.; Hripcsak, G.; et al. Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu. Symp. Proc.* **2014**, *2014*, 907–916. [PubMed]
7. Teixeira, P.L.; Wei, W.-Q.; Cronin, R.M.; Mo, H.; VanHouten, J.P.; Carroll, R.J.; LaRose, E.; Bastarache, L.A.; Rosenbloom, S.T.; Edwards, T.L.; et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J. Am. Med. Inform. Assoc.* **2016**, *24*, 162–171. [CrossRef] [PubMed]
8. McDonough, C.W.; Babcock, K.; Chucri, K.; Crawford, D.C.; Bian, J.; Modave, F.; Cooper-DeHoff, R.M.; Hogan, W.R. Optimizing identification of resistant hypertension: Computable phenotype development and validation. *Pharmacoepidemiol. Drug Saf.* **2020**, *29*, 1393–1401. [CrossRef] [PubMed]
9. Tan, C.C.; Lam, C.S.P.; Matchar, D.B.; Zee, Y.K.; Wong, J.E.L. Singapore's health-care system: Key features, challenges, and shifts. *Lancet* **2021**, *398*, 1091–1104. [CrossRef] [PubMed]
10. MOH Clinical Practice Guidelines on Hypertension. Ministry of Health, Singapore. 2023. Available online: https://www.moh.gov.sg/hpp/doctors/guidelines/GuidelineDetails/cpgmed_hypertension (accessed on 16 February 2023).
11. Parikh, R.; Mathai, A.; Parikh, S.; Chandra Sekhar, G.; Thomas, R. Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.* **2008**, *56*, 45–50. [CrossRef]
12. Leeflang, M.M.; Rutjes, A.W.; Reitsma, J.B.; Hooft, L.; Bossuyt, P.M. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ* **2013**, *185*, E537–E544. [CrossRef] [PubMed]

13. Huang, W.; Lee, S.G.S.; How, C.H. Management of the heart failure patient in the primary care setting. *Singapore Med. J.* **2020**, *61*, 225–229. [CrossRef]
14. Whelton, P.K.; Carey, R.M.; Aronow, W.S.; Casey, D.E., Jr.; Collins, K.J.; Himmelfarb, C.D.; DePalma, S.M.; Gidding, S.; Jamerson, K.A.; Jones, D.W.; et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension* **2018**, *71*, e13–e115. [CrossRef]