BMJ Open Diabetes Research & Care

Application of machine learning techniques to understand ethnic differences and risk factors for incident chronic kidney disease in Asians

Cynthia Ciwei Lim ,[1] Feng He,[2] Jialiang Li,[3] Yih Chung Tham,[2,4] Chieh Suai Tan,[1] Ching-Yu Cheng,[2,4] Tien-Yin Wong,[2,4,5] Charumathi Sabanayagam [2,4,5]

To cite: Lim CC, He F, Li J, *et al*. Application of machine learning techniques to understand ethnic differences and risk factors for incident chronic kidney disease in Asians. *BMJ Open Diab Res Care* 2021;**9**:e002364. doi:10.1136/bmjdrc-2021-002364

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjdrc-2021-002364).

Received 3 May 2021
Accepted 14 November 2021

For numbered affiliations see end of article.

Correspondence to
Dr Charumathi Sabanayagam; charumathi.sabanayagam@seri.com.sg

ABSTRACT

Introduction Chronic kidney disease (CKD) is increasing in Asia, but there are sparse data on incident CKD among different ethnic groups. We aimed to describe the incidence and risk factors associated with CKD in the three major ethnic groups in Asia: Chinese, Malays and Indians.

Research design and methods Prospective cohort study of 5580 general population participants age 40–80 years (2234 Chinese, 1474 Malays and 1872 Indians) who completed both baseline and 6-year follow-up visits. Incident CKD was defined as an estimated glomerular filtration rate (eGFR) <60 mL/min/1.73 m^2 in those free of CKD at baseline.

Results The 6-year incidence of CKD was highest among Malays (10.0%), followed by Chinese (6.1%) and Indians (5.8%). Logistic regression showed that older age, diabetes, higher systolic blood pressure and lower eGFR were independently associated with incident CKD in all three ethnic groups, while hypertension and cardiovascular disease were independently associated with incident CKD only in Malays. The same factors were identified by machine learning approaches, gradient boosted machine and random forest to be the most important for incident CKD. Adjustment for clinical and socioeconomic factors reduced the excess incidence in Malays by 60% compared with Chinese but only 13% compared with Indians.

Conclusion Incidence of CKD is high among the main Asian ethnic groups in Singapore, ranging between 6% and 10% over 6 years; differences were partially explained by clinical and socioeconomic factors.

Significance of this study

What is already known about this subject?

- Aging populations and greater prevalence of metabolic risk factors such as diabetes and hypertension have contributed to increased CKD but it is unknown if there are disparities in the risk and contributory factors for incident chronic kidney disease (CKD) in the major ethnic groups in Asia.

What are the new findings?

- The 6-year incidence of CKD in Chinese, Malays and Indians was 6.1%, 10% and 5.8%, respectively. Older age, diabetes, higher systolic blood pressure and lower eGFR were independently associated with incident CKD in all ethnicities, while hypertension and cardiovascular disease were associated with incident CKD only in Malays.

How might these results change the focus of research or clinical practice?

- Significant ethnic disparities in incident CKD in Asians were partially explained by clinical and socioeconomic factors that can be targeted to reduce incident CKD.

INTRODUCTION

Chronic kidney disease (CKD) is recognized to be a global health burden.[1] Aging populations and greater prevalence of metabolic risk factors such as diabetes and hypertension have contributed to increased CKD worldwide and especially in Asia.[2 3] In the Global Burden of Disease Study 2017,[2] global life expectancy increased by 7.4 years from 65.6 years in 1990 to 73.0 years in 2017. CKD contributed to one of the largest increases in disability-adjusted life years,[2] and CKD-related deaths increased by 40% among those aged 50–69 years and by 42% among those 70 years and older.[4] CKD is also costly to patients and the society.[5] The median values of total direct and out-of-pocket healthcare expenditures were $12 877 and $1439, respectively, among individuals with CKD in the USA,[5] more than five times the expenditures of those without CKD. Hence, there is a need to establish incidence of CKD in the general population to better anticipate and prepare for the challenges that CKD brings to the healthcare system. Annualized incident CKD rates among the general population are estimated to be 0.7%–1.2% in North America and Europe,[6–9] with fewer studies in Asia. While some East Asian countries such as Taiwan, Korea and Japan have reported rates of 0.9%–2%,[10–13] there are

scant data in other ethnic groups such as Malays. Other than a hospital-based study of type 2 diabetes that evaluated CKD progression,[14] data on incident CKD in Singapore are sparse.

An earlier cross-sectional study reported CKD prevalence (higher in Malays and Indians) to be different among the three ethnic groups in Singapore.[15] Ethnic disparities (eg, white vs Hispanic and black people) in CKD prevalence have also been reported in North America,[16] where they were attributed to genetic differences and/or socioeconomic barriers to accessing healthcare.[16 17] However, CKD risk and its contributory factors appear to differ between Asian and Caucasian populations so previous studies may not be generalizable.[18] There is growing recognition that in order to address these disparities in health outcomes, there is a need for a culturally competent healthcare system that first acknowledges the differences and the contributory reasons and then adapts services to meet the unique needs of the population.[17] To address these gaps, we aimed to describe and compare the incidence and factors associated with incident CKD in the three major ethnic groups in Asia and Singapore: Chinese, Malays and Indians. Furthermore, to better understand risk factors for incident CKD, and possible ethnic differences, we applied machine learning techniques in addition to standard logistic regression (LR) models.

RESEARCH DESIGN AND METHODS

Study population

The Singapore Epidemiology of Eye Diseases (SEED) Study is a large population-based prospective cohort study of Chinese, Malay and Indian adults aged 40–80 years at baseline.[19] Three independent studies, the Singapore Malay Eye Study (2004–2006), the Singapore Indian Eye Study (2007–2009) and the Singapore Chinese Eye Study (2009–2011) conducted by the Singapore Eye Research Institute were combined. Detailed methodology for these studies was previously reported.[20–22] In brief, age-stratified random sampling from computer-generated random lists of individuals 40–80 years of age residing in the same geographical area in Singapore generated a sampling frame of 6350 Chinese, 5600 Malays and 6350 Indians. A total of 10 033 participants comprising 3353 Chinese, 3280 Malays and 3400 Indians participated in the baseline visit and 6762 (78.8%) returned for the follow- up visit.[19] For this study, we included participants who attended both baseline and 6-year follow-up visits. After excluding those with missing values on estimated glomerular filtration rate (eGFR) at baseline or follow-up (n=597), those with prevalent CKD at baseline (n=524) and those with missing data on key covariates including hypertension, body mass index (BMI), lipid profile, current smoking status, alcohol consumption and education category (n=61), 5580 SEED participants were included for the current analysis. Figure 1 shows the selection of the SEED participants included in the analysis.

Data collection

An interviewer-administered questionnaire was used to collect participants' sociodemographic (age, gender), socioeconomic (highest education attained), lifestyle (current smoking, alcohol consumption) and medical history as previously described.[23] Physical examination included height, weight and blood pressure (BP) measurements. We calculated BMI as weight in kilograms divided by height in meters squared. Obesity was defined as BMI ≥25 kg/m^2. Hypertension was defined in the presence of systolic BP ≥140 mm Hg, diastolic BP ≥90 mm Hg; participants reported hypertension diagnosed by physicians or use of BP lowering therapy. Among those with hypertension, BP control was defined as having BP <140/90 mm Hg.[24] Diabetes mellitus was defined as random serum glucose level ≥11.1 mmol/L, glycosylated hemoglobin (HbA1c)≥6.5%; participants reported diabetes diagnosed by physicians or use of glucose-lowering treatment.[25] Among those with diabetes, glycemic control was defined as HbA1c <7%.[26] Cardiovascular disease was defined as self-reported myocardial infarction, angina, or stroke. Non-fasting serum lipid, glucose, HbA1c and creatinine were evaluated. Dyslipidemia was defined as total cholesterol ≥6.2 mmol/L, low-density lipoprotein cholesterol ≥4.1 mmol/L, and high-density lipoprotein cholesterol <1.0 mmol/L; self-reported physician-diagnosed dyslipidemia; or use of statin medication. Serum creatinine was measured using an enzymatic method calibrated to the National Institute of Standard and Technology liquid chromatography isotope dilution mass spectrometry method.[19] eGFR was calculated using the CKD Epidemiology Collaboration (CKD-EPI) equation.[27] Laboratory investigations were conducted at hospitals accredited by the College of American Pathologists.

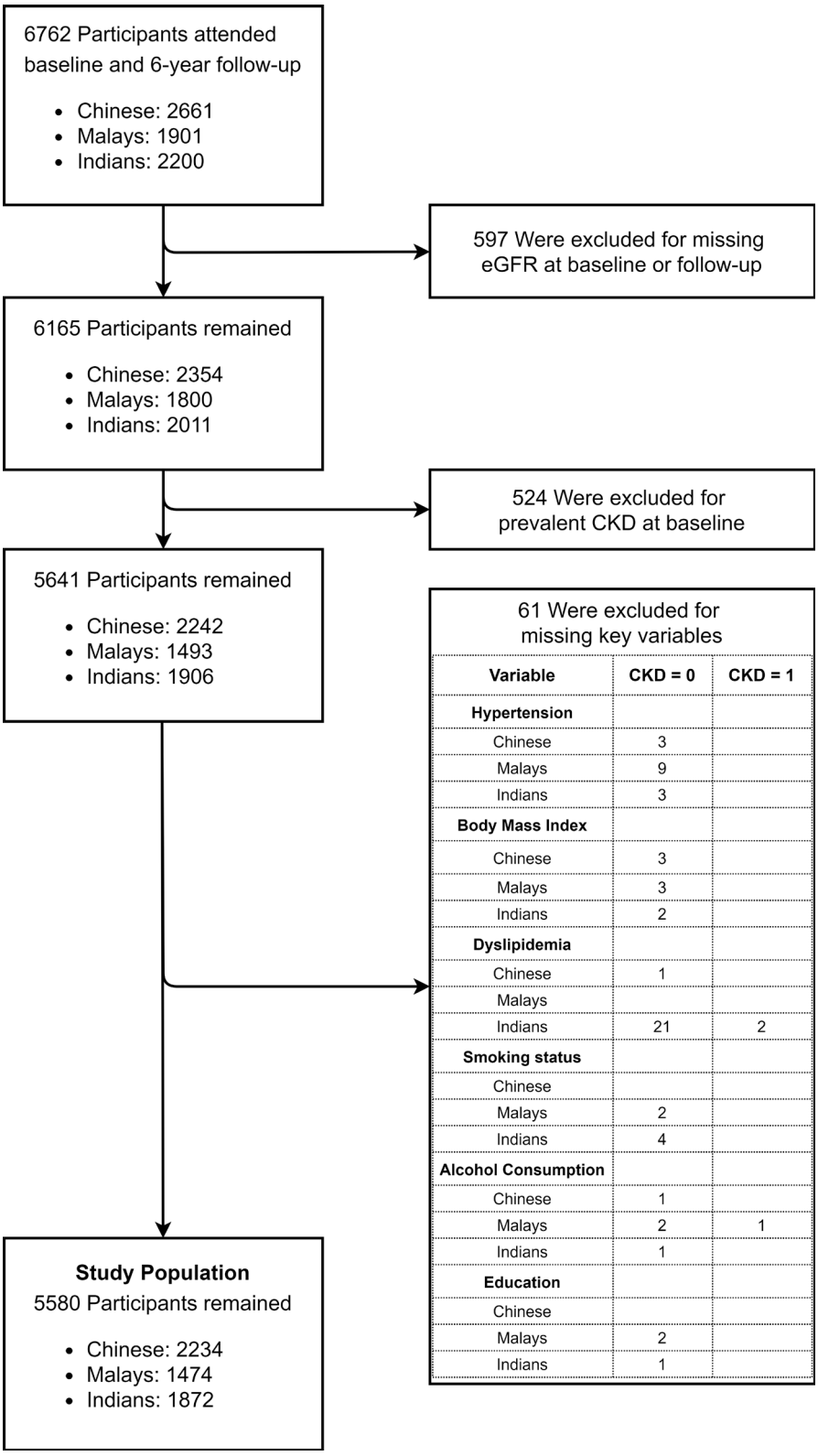
Participants gave written informed consent before enrolment.

Outcome definition

Incident CKD was defined when individuals with eGFR ≥60 mL/min/1.73 m^2 at enrollment subsequently had eGFR <60 mL/min/1.73 m^2 at follow-up. The reduction in eGFR at follow-up was calculated as a percentage of the baseline eGFR at enrollment, that is, ((eGFR at baseline – eGFR at follow-up) / eGFR at baseline) *100%.

Statistical analysis

Baseline characteristics by ethnicity and incident CKD status were examined using means (SD), median (IQR) or count (percentage) and compared using Mann-Whitney test (ie, two-sample Wilcoxon test) or Fisher's exact test as appropriate for the variable. Sociodemographic and clinical characteristics by ethnicity at baseline and follow-up were compared using Kruskal-Wallis rank sum test or χ^2 test as appropriate for the variable. LR was used to calculate the age-adjusted and sex-adjusted and multivariable-adjusted ORs and 95% CIs for factors associated with incident CKD in each ethnic group, while



Variable	CKD = 0	CKD = 1
Hypertension		
Chinese	3	
Malays	9	
Indians	3	
Body Mass Index		
Chinese	3	
Malays	3	
Indians	2	
Dyslipidemia		
Chinese	1	
Malays		
Indians	21	2
Smoking status		
Chinese		
Malays	2	
Indians	4	
Alcohol Consumption		
Chinese	1	
Malays	2	1
Indians	1	
Education		
Chinese		
Malays	2	
Indians	1	

Figure 1 Flow diagram of participant exclusion. CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate.

linear regression was used to evaluate factors associated with the continuous outcome of percentage reduction in eGFR. Covariates were selected based on established prognostic factors according to known literature.[28] To account for attrition bias, we performed a supplementary analysis using inverse probability weighting (IPW) and obtained weighted regression coefficients for comparison with the unweighted ones. Statistical significance was defined to be two-sided p values <0.05. To further validate the findings in LR and to evaluate the importance of each risk factor for incident CKD, we employed two classic machine learning approaches, gradient boosted machine

(GBM),[29] and random forest (RF).[30] In GBM, the relative influence score measures the proportional contribution of a variable on the model performance, with all scores sum up to 100%.[29] In RF, the mean decrease in accuracy measures the change in the prediction accuracy resulted from the exclusion (or permutation) of a variable.[30]

To evaluate the extent that clinical, metabolic (cardiovascular disease, dyslipidemia, diabetes, hypertension, systolic BP), socioeconomic (education) and behavioral (smoking, obesity, diabetes control, BP control) factors may account for the excess CKD risk in the Malay cohort, we calculated the reduction in ORs associated with adjustment for these factors using the formula[15]

$$\frac{OR_1 - OR_i}{OR_1 - 1},\ i = 2, 3, 4$$

where OR_1 is the OR of incident CKD in Malays versus Chinese and Malays versus Indians, adjusted for age and sex only (model 1), and OR_i is the OR of further adjusted models 2 (model 1 with additional clinical and metabolic factors), 3 (model 1 with additional socioeconomic and behavioral factors) and 4 (included all factors).

Age-standardized prevalence of risk factors and age-standardized CKD incidence were estimated using the population distribution of the 2010 Singapore Census (only included Chinese, Malays, and Indians, who were Singapore citizens or permanent residents of age 40–80 years). Annual incidence was calculated by dividing the cumulative incidence by the summed person-years.

To establish the proportion of all cases of incident CKD in the total population that could be attributed to the exposure to the binary risk factors that were significant in the multivariable model, we estimated the population attributable risks (PARs) due to hypertension, diabetes and cardiovascular disease using Levin's formula:

$$\text{PAR}\% = \frac{\text{Prevalence} \times (\text{Relative Risk} - 1) \times 100}{[\text{Prevalence} \times (\text{Relative Risk} - 1) + 1]}$$

where the relative risk was estimated by the adjusted OR.[31]

All analyses were performed using R V.4.0.0 (R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/).

RESULTS

We identified 5580 individuals (1474 Malays, 2234 Chinese and 1872 Indians) with eGFR ≥60 mL/min/1.73 m^2 at baseline. The mean eGFR values at baseline were lower in Malays (82.9 mL/min/1.73 m^2) compared with Chinese and Indians (92.8 mL/min/1.73 m^2 and 91.5 mL/min/1.73 m^2, respectively). The median follow-up was 6.7 (5.7–7.3) years in Malays, 6.0 (5.3–6.5) years in Chinese and 5.9 (5.5–6.6) years in Indians. Metabolic risk factors such as diabetes, hypertension and dyslipidemia were more frequent at follow-up than at baseline in all ethnic groups (online supplemental etable 1), but the relative distributions among the ethnic groups remained similar. At both baseline and follow-up, Malays were more likely to be current smokers, have obesity, hypertension with higher systolic and diastolic BP and lower eGFR compared with other groups, while Indians were more likely to have diabetes, dyslipidemia and have higher glucose levels than other ethnic groups. Among those with diabetes at baseline, Malays were less likely to have adequate diabetic control and have antidiabetic medications, compared with the other ethnic groups. Among those with hypertension at baseline, Malays were less likely to have BP control and have antihypertensive medications including ACE inhibitors or angiotensin II receptor blockers, compared with other ethnic groups.

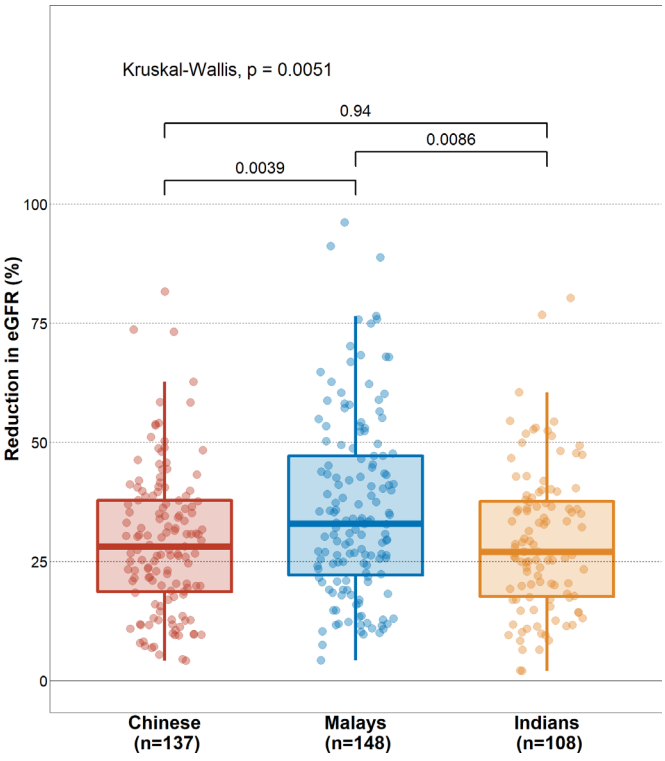


Figure 2 Median reduction in estimated glomerular filtration rate (eGFR) was significantly greater in Malays (33.0%, IQR 22.1%–47.2%) compared with Chinese (28.2%, IQR 18.7%–37.9%) or Indians (27.0%, IQR 17.6%–37.6%).

The 6 year incidence of CKD was higher in Malays (10.0%) followed by Chinese (6.1%) and Indians (5.8%). Consequently, the age-standardized annual incidence was significantly higher in Malays and lower in Chinese and Indians (online supplemental etable 2). Incident CKD in Malays was more severe, with reduced eGFR <30 mL/min/1.73 m^2 in 1.2% compared with 0.2% in Chinese and 0.2% in Indians. Figure 2 shows that the median reduction in eGFR was significantly greater in Malays (33.0%, IQR 22.1%–47.2%) compared with Chinese (28.2%, IQR 18.7%–37.9%, p value=0.004) or Indians (27.0%, IQR 17.6%–37.6%, p value=0.009).

Table 1 shows the clinical characteristics at baseline stratified by incident CKD in each ethnic group. Compared with those without incidence CKD, those with incident CKD were older, had lower education level and eGFR

Table 1 Baseline characteristics of SEED participants stratified by ethnicity and incident CKD status

Variable	Chinese			Malays			Indians		
	No CKD (n=2097)	CKD (n=137)	P value*	No CKD (n=1326)	CKD (n=148)	P value*	No CKD (n=1764)	CKD (n=108)	P value*
Age, year	56.7 (8.2)	67.1 (8)	<0.001	54.3 (9.1)	63.2 (8)	<0.001	54.8 (8.2)	65.7 (8)	<0.001
Gender, N (%)			<0.001			0.118			0.692
Female	1097 (95.8)	48 (4.2)		735 (91.1)	72 (8.9)		875 (94.5)	51 (5.5)	
Male	1000 (91.8)	89 (8.2)		591 (88.6)	76 (11.4)		889 (94)	57 (6)	
Education†, N (%)			<0.001			<0.001			<0.001
Primary/below	966 (91.5)	90 (8.5)		806 (87.4)	116 (12.6)		861 (92.3)	72 (7.7)	
Secondary/above	1131 (96)	47 (4)		520 (94.2)	32 (5.8)		903 (96.2)	36 (3.8)	
Diabetes mellitus, N (%)			<0.001			<0.001			<0.001
No	1820 (95.8)	79 (4.2)		1028 (93.6)	70 (6.4)		1209 (97.1)	36 (2.9)	
Yes	277 (82.7)	58 (17.3)		298 (79.3)	78 (20.7)		555 (88.5)	72 (11.5)	
Hypertension, N (%)			<0.001			<0.001			<0.001
No	1018 (98.3)	18 (1.7)		572 (98.3)	10 (1.7)		862 (98.4)	14 (1.6)	
Yes	1079 (90.1)	119 (9.9)		754 (84.5)	138 (15.5)		902 (90.6)	94 (9.4)	
Current smoking, N (%)			0.694			0.585			0.474
No	1828 (93.9)	118 (6.1)		1064 (89.7)	122 (10.3)		1516 (94)	96 (6)	
Yes	269 (93.4)	19 (6.6)		262 (91)	26 (9)		248 (95.4)	12 (4.6)	
Dyslipidemia, N (%)			<0.001			0.020			<0.001
No	1206 (95.9)	52 (4.1)		843 (91.4)	79 (8.6)		1005 (96.3)	39 (3.7)	
Yes	891 (91.3)	85 (8.7)		483 (87.5)	69 (12.5)		759 (91.7)	69 (8.3)	
Cardiovascular disease, N (%)			<0.001			0.002			<0.001
No	2008 (94.6)	115 (5.4)		1245 (90.7)	128 (9.3)		1594 (95.2)	81 (4.8)	
Yes	89 (80.2)	22 (19.8)		81 (80.2)	20 (19.8)		170 (86.3)	27 (13.7)	
Obesity, N (%)			0.003			0.064			1.000
No	1456 (94.9)	78 (5.1)		546 (91.8)	49 (8.2)		760 (94.3)	46 (5.7)	
Yes	641 (91.6)	59 (8.4)		780 (88.7)	99 (11.3)		1004 (94.2)	62 (5.8)	
Body mass index, kg/m^2	23.7 (3.6)	24.7 (3.2)	<0.001	26.4 (4.9)	27.6 (4.5)	0.001	26.2 (4.4)	26.3 (4.3)	0.955
Blood glucose, mmol/L	6.1 (2.2)	7.8 (3.9)	<0.001	6.2 (3)	8.7 (5.5)	<0.001	6.8 (3.1)	7.8 (3.3)	<0.001
Systolic blood pressure, mm Hg	133.5 (18.2)	145.6 (17.8)	<0.001	140.2 (21)	158.5 (22.3)	<0.001	132.4 (18.6)	147.8 (19.6)	<0.001
Diastolic blood pressure, mm Hg	77.6 (9.8)	79 (9.4)	0.112	78.9 (10.5)	81.6 (12.8)	0.029	78.1 (10.1)	78.2 (9.9)	0.913
Pulse pressure, mm Hg	55.9 (14.3)	66.6 (15.7)	<0.001	61.3 (16)	76.9 (16.7)	<0.001	54.4 (14.3)	69.6 (16.5)	<0.001
Antihypertensive medication use‡, N (%)			<0.001			<0.001			<0.001

Continued

Table 1 Continued

Variable	Chinese			Malays			Indians		
	No CKD (n=2097)	CKD (n=137)	P value*	No CKD (n=1326)	CKD (n=148)	P value*	No CKD (n=1764)	CKD (n=108)	P value*
No	513 (94.3)	31 (5.7)		482 (87.8)	67 (12.2)		375 (95.4)	18 (4.6)	
Yes	566 (86.5)	88 (13.5)		272 (79.3)	71 (20.7)		527 (87.4)	76 (12.6)	
Total cholesterol, mmol/L	5.5 (1)	5.2 (1.1)	<0.001	5.6 (1.1)	5.5 (1.2)	0.734	5.3 (1.1)	4.9 (1.1)	0.005
HDL cholesterol, mmol/L	1.3 (0.4)	1.2 (0.4)	0.002	1.4 (0.3)	1.3 (0.3)	<0.001	1.1 (0.3)	1.1 (0.3)	0.512
Baseline eGFR, mL/min/1.73 m^2	94.1 (12.7)	73.4 (10.3)	<0.001	83.9 (14)	74.4 (11.8)	<0.001	92.6 (12.8)	73.7 (10.3)	<0.001
Alcohol consumption, N (%)			0.215			0.507			0.192
No	1850 (93.6)	126 (6.4)		1301 (89.8)	147 (10.2)		1531 (94.5)	89 (5.5)	
Yes	247 (95.7)	11 (4.3)		25 (96.2)	1 (3.8)		233 (92.5)	19 (7.5)	

Data presented are counts (row percentage) or means (SD).
*P value represents difference in characteristics by incident CKD status based on Mann-Whitney test or Fisher's exact test as appropriate for the variable.
†Highest education level attained was categorized as no formal education or completed primary school education versus completed at least secondary school with Singapore Cambridge General Certificate of Education Ordinary level or equivalent.
‡Among those with self-reported hypertension.
CKD, chronic kidney disease; eGFR, estimated glomerular filtration rate; HDL, high-density lipoprotein; SEED, Singapore Epidemiology of Eye Disease.

but higher blood glucose, systolic BP and pulse pressure. Diabetes, hypertension, cardiovascular disease, were more frequent in those with incident CKD in all three ethnic groups. Chinese and Malays with incident CKD were more likely to be male, obese and had higher BMI and lower high-density lipoprotein (HDL)-cholesterol than those without incident CKD. Chinese and Indians with incident CKD were more likely to have dyslipidemia and lower total cholesterol than those without incident CKD. In LR models stratified by ethnicity (table 2), older age, diabetes, higher systolic BP and lower eGFR were independently associated with incident CKD in all three ethnic groups, while hypertension and cardiovascular disease were independently associated with incident CKD only in Malays. As 'hypertension' was a broad categorical variable, there may be residual confounding by BP. Online supplemental etable 3 shows that the exclusion of systolic BP from the LR model resulted in higher adjusted ORs for hypertension in all ethnicities, while the adjusted ORs for the other predictors were similar to the LR model that included systolic BP. In the linear regression model (online supplemental etable 4), older age, diabetes, higher systolic BP and lower eGFR remained consistently associated with greater reduction in eGFR in all three ethnic groups. The magnitude of the association between systolic BP and percentage reduction in eGFR was largest in Malays among the ethnic groups. In addition, cardiovascular disease was independently associated with greater percentage reduction in eGFR in all three groups, while male gender was significantly associated with greater percentage reduction in eGFR among Chinese and Malays. Supplementary analysis with IPW identified the same factors with similar risk estimates for both incident CKD and eGFR reduction (data not shown). Both GBM and RF identified eGFR, age, systolic BP and diabetes to be the most important variables in incident CKD prediction (online supplemental efigure 1). GBM also found hypertension and cardiovascular disease to be influential in Malays.

The estimated PAR of diabetes for incident CKD was highest among Indians (45.2%) compared with Malays (35.4%) and Chinese (33.2%) based on age-standardized prevalences among those aged 40–80 years (online supplemental etable 5). Among Malays, hypertension had higher PAR (54.7%) than diabetes and cardiovascular disease (7.5%). Table 3 shows that the odds of incident CKD were markedly attenuated after adjustment for clinical, metabolic, socioeconomic and behavioural factors when comparing Malays and Chinese, and to a lesser extent when comparing Malays and Indians. Adjustment for all factors reduced the excess incidence in Malays by 64% compared with Chinese but only 19% compared with Indians.

In the subgroup of 1338 individuals with diabetes, incident diabetic CKD occurred in 208 (15.5%). Like the main analysis, incident diabetic CKD was most frequent in Malays (20.7%), compared with Chinese (17.3%) and Indians with diabetes (11.5%). Online supplemental

Table 2 Multivariable predictors of incident CKD by ethnicity (n=5580)

Variable	Chinese, n=2234 Incident CKD, n=137 (6.1%)		Malays, n=1474 Incident CKD, n=148 (10.0%)		Indians, n=1872 Incident CKD, n=108 (5.8%)	
	Age, sex-adjusted OR (95% CI)	Multivariable OR (95% CI)	Age, sex-adjusted OR (95% CI)	Multivariable OR (95% CI)	Age, sex-adjusted OR (95% CI)	Multivariable OR (95% CI)
Age, per 10 years increase	3.88 (3.11 to 4.84)	1.68 (1.26 to 2.25)	2.71 (2.24 to 3.29)	1.95 (1.53 to 2.49)	3.75 (2.97 to 4.73)	1.75 (1.28 to 2.38)
Gender, male versus female	2.19 (1.49 to 3.21)	1.12 (0.68 to 1.82)	1.26 (0.88 to 1.79)	1.00 (0.64 to 1.56)	1.03 (0.68 to 1.56)	0.68 (0.41 to 1.14)
Secondary and above education versus primary and below	0.81 (0.54 to 1.21)	0.78 (0.49 to 1.22)	0.85 (0.53 to 1.34)	1.05 (0.64 to 1.71)	0.79 (0.5 to 1.24)	0.84 (0.50 to 1.39)
Current smoking, Yes versus No	1.02 (0.58 to 1.78)	1.11 (0.60 to 2.07)	1.01 (0.60 to 1.69)	1.45 (0.82 to 2.54)	1.29 (0.65 to 2.56)	1.60 (0.73 to 3.51)
Diabetes, Yes versus No	3.44 (2.33 to 5.07)	4.56 (2.84 to 7.31)	3.20 (2.22 to 4.61)	3.32 (2.23 to 4.93)	3.00 (1.94 to 4.63)	3.65 (2.20 to 6.05)
Hypertension, Yes versus No	2.95 (1.74 to 5.00)	1.26 (0.65 to 2.42)	6.37 (3.27 to 12.4)	3.13 (1.49 to 6.57)	2.96 (1.62 to 5.39)	1.41 (0.68 to 2.92)
Dyslipidemia, Yes versus No	1.64 (1.12 to 2.41)	0.84 (0.53 to 1.32)	1.17 (0.81 to 1.68)	0.77 (0.51 to 1.15)	1.41 (0.92 to 2.17)	0.94 (0.56 to 1.56)
Systolic blood pressure, per SD increase	1.30 (1.07 to 1.57)	1.31 (1.02 to 1.69)	1.73 (1.45 to 2.07)	1.46 (1.18 to 1.81)	1.62 (1.32 to 1.98)	1.55 (1.20 to 2.00)
Estimated glomerular filtration rate, per unit decrease	1.12 (1.09 to 1.14)	1.13 (1.10 to 1.15)	1.04 (1.03 to 1.06)	1.05 (1.03 to 1.07)	1.11 (1.09 to 1.14)	1.12 (1.10 to 1.15)
Obesity, Yes versus No	2.10 (1.44 to 3.07)	1.38 (0.89 to 2.16)	1.69 (1.15 to 2.48)	1.30 (0.86 to 1.97)	1.37 (0.89 to 2.09)	0.85 (0.52 to 1.40)
Cardiovascular disease, Yes versus No	2.16 (1.23 to 3.79)	1.54 (0.80 to 2.96)	1.97 (1.12 to 3.44)	2.26 (1.22 to 4.17)	1.83 (1.1 to 3.05)	1.35 (0.74 to 2.45)

Multivariable model adjusted for age, gender, education, current smoking, diabetes, dyslipidemia, systolic blood pressure, estimated glomerular filtration rate, obesity and cardiovascular disease.
CKD, chronic kidney disease.

etable 6 shows that lower eGFR was an independent predictor for incident diabetic CKD in all ethnic groups. Additionally, older age, hypertension, higher systolic BP, HbA1c and diabetes duration predicted incident diabetic CKD among Malays with diabetes, while higher systolic BP and HbA1c predicted incident diabetic CKD in Indians with diabetes.

DISCUSSION

In this prospective study of 5580 multiethnic Asians in the general population with a median follow-up of 6.1 years, incident CKD was more severe and more frequent in Malays compared with Chinese and Indians. Older age, diabetes, higher systolic BP and lower eGFR were independently associated with incident CKD in all three ethnic groups, while hypertension and cardiovascular disease were independently associated with incident CKD only in Malays. The estimated PAR of diabetes for incident CKD was 45.2% among Indians; while the PAR of hypertension was 54.7% among Malays. Adjustment for clinical, metabolic, socioeconomic and behavioural factors reduced the excess risk in Malays by 64% compared with Chinese but only 19% compared with Indians.

The annualized incident CKD rate was highest among Malays (1.3%) but the rates for all groups were similar to annualized rates of 0.9%–2% reported by general population studies in Taiwan, Korea and Japan.[10–13] While population-based data from other ethnicities such as Malays and Indians are sparse, a retrospective cohort study of 460 individuals (25.4% Malays, 50.0% Chinese and 23.5% Indians) with hypertension from a Malaysian university medical centre's primary care clinic reported that the incidence of CKD was 30.9% over 10 years.[32] Our findings of estimated crude annual incidence and age-standardized annual incidence for each ethnicity are useful in informing the burden of incident CKD in the general population, since the estimated incidence rate of 1%–1.3% per year would translate to 44 000 incident CKD among the residential adult population of 3.2 million.[33] Ethnic disparities in CKD prevalence were observed in an earlier, separate cohort of multiethnic general population study.[15] While few Asian studies have compared incident CKD by ethnicity, disparities in kidney disease have been observed in North America where incident CKD varied by Hispanic/Latino heritage,[34] and African-Americans suffer disproportionately from kidney disease.[16] Measures of socioeconomic status attenuated the relation between African-American ethnicity and CKD but did not eliminate them.[16] Likewise, this study found that adjustment for clinical, metabolic, socioeconomic and behavioural factors only partially explained the excess risk for incident CKD in Malays when compared with Chinese or Indians. The remaining excess risk unexplained by the multivariable model may be related to residual confounding from variables not included in this study, including other social determinants of health and genomic differences.[16 35] Prior studies have noted ethnic differences in health literacy

Table 3 Factors affecting the excess incidence of CKD in Malays and Indians compared with Chinese

	Malays versus Chinese		Malays versus Indians	
Adjustment models	OR (95% CI)	% Reduction in excess risk*	0R (95% CI)	% Reduction in excess risk*
Age and sex (1)†	2.22 (1.71 to 2.88)	Reference	1.90 (1.45 to 2.50)	Reference
Clinical and metabolic factors (2)‡	1.64 (1.24 to 2.17)	48	1.85 (1.38 to 2.50)	5
Socioeconomic and behavioral factors (3)§	1.51 (1.14 to 2.00)	58	1.67 (1.25 to 2.24)	25
Fully adjusted (4)¶	1.44 (1.08 to 1.94)	64	1.73 (1.27 to 2.35)	19

*Per cent reduction in incidence difference defined by the formula:[15] $\frac{OR_1 - OR_i}{OR_1 - 1}$, $i = 2, 3, 4$ where OR_1 is the OR of incident CKD in Malays versus Chinese and Malays versus Indians, adjusted for age and sex only (model 1), and OR_i is the OR after adjustment for variables in models 2, 3, 4.
†Model 1, ORs (95% CI) of incident CKD in association with ethnicity adjusting for age and sex.
‡Model 2, adjusted for variables in model 1 plus diabetes, hypertension, systolic BP, history of cardiovascular disease, and dyslipidemia.
§Model 3, adjusted for variables in model 1 plus education, current smoking, obesity, diabetes control (HbA1c<7%) and BP control (systolic BP <140 mm Hg and diastolic BP <90 mm Hg).
¶Model 4 adjusted for all variables in models 1 to 3.
BP, blood pressure; CKD, chronic kidney disease; HbA1c, glycosylated hemoglobin.

and health information-seeking behaviours possibly related to language barriers or cultural norms,[36 37] which in turn may translate to differences in risk factor awareness and control shown in our study and highlighted in others.[38 39] Since interventions improved outcomes in those with low health literacy,[40] targeted strategies such as patient education and health policy change will be required to reduce incident CKD.[41]

Diabetes causes microvascular disease that leads to glomerular hyperfiltration with subsequent glomerulosclerosis, tubulointerstitial inflammation and fibrosis and is an established risk factor for progressive kidney disease.[23 35] Thus, it was unsurprising that incident CKD among diabetes was twofold or threefold that of the general cohort in all ethnic groups. Similarly, an analysis of 34 international cohorts from the CKD Prognosis Consortium reported incident CKD in 14.9% of over 4 million participants without diabetes during a mean follow-up of 4.2 years and 40% of 781 627 participants with diabetes during a mean follow-up of 3.9 years.[28] In our study, older age, higher systolic BP and lower eGFR were also independently associated with incident CKD in all three ethnic groups. These results were consistently found in traditional LR and both machine learning models. Hypertension and cardiovascular disease were identified to be important variables for incident CKD in Malays in both LR and GBM, but not in RF. The machine learning models complement LR, which assesses the association in a unit-dependent manner but fails to address the difference in the variable range or category. Instead, both GBM and RF calculate the effect of a variable as its overall contribution to the model performance, which is unit-free and applicable to various variable ranges or categories. For example, our LR model showed eGFR, age, systolic BP, and diabetes to be significant for CKD prediction, but it was machine learning that identified eGFR as the most influential variable of the four. Hence machine learning provided a simple and intuitive measure for the comparison of CKD risk factors, which is not directly achievable in LR where the OR and the p value need to be considered simultaneously. While machine learning methods can capture non-linear relationships and interactions in the variables,[30 42] their performance for disease risk modelling may not be superior to traditional LR,[43] especially when the variables are few and the sample size is small.[44] Using traditional LR, the CKD Prognosis Consortium similarly found that among participants with no diabetes, older age, lower eGFR, hypertension and cardiovascular disease were associated with increased risk of incident CKD.[28] Other risk factors identified by the Consortium but not significantly associated with incident CKD in our study were female gender, ever-smoker and BMI.[28] While obesity was associated with incident CKD in Chinese in the univariate analysis, the association was lost after adjusting for all other factors. In contrast, a systematic review of 39 cohorts that included 630 677 participants with a mean follow-up of 6.8 years found that incident CKD was increased in obesity (pooled relative risk 1.28, 95% CI 1.07 to 1.54).[45] Lower eGFR was an independent predictor for incident diabetic CKD in all ethnic groups. Additionally, older age, hypertension, higher systolic BP, HbA1c and diabetes duration predicted incident diabetic CKD among Malays with diabetes, while higher systolic BP and HbA1c predicted incident diabetic CKD in Indians with diabetes. These were similar to findings by the CKD Prognosis Consortium.[28] Other risk factors identified by the Consortium but not significantly associated with incident diabetic CKD in our study were female gender, cardiovascular disease and BMI.[28]

There are some limitations in this study. CKD was defined based on eGFR, similar to the majority of studies on incident CKD,[45] while albuminuria was not included in the definition of baseline or incident CKD since urine albuminuria was available only in a third of the Malay participants at baseline

(those with known diabetes and one in five with no diabetes) and not quantified during follow-up. Incident CKD was defined using a single laboratory measurement and may overestimate incident CKD without a repeat measurement 3 months apart. However, the aforementioned systematic review noted no difference in the comparison between studies with and without repeated measurements of serum creatinine.[45] Antihypertensive medication type, dose and duration were not evaluated as factors for the outcome since information on the type of antihypertensive was incomplete while the dose and duration were not assessed. As this study necessarily included only participants who attended and had renal function tests at both baseline and follow-up visits, there may be loss of follow-up and survival bias. As this study included older individuals 40–80 years old, some individuals may have died or developed significant disability that led to non-attendance at the follow-up visit. Compared with individuals who did not return for the follow-up visit, participants who returned for the follow-up visit were younger, more likely to be female, Chinese, attained secondary school or above education, and less likely to be Malay or Indian, have diabetes, hypertension, cardiovascular disease, current smoking (online supplemental etable 7). They also had lower systolic and diastolic BP, glucose, glycated hemoglobin and higher eGFR. Thus both loss to follow-up and survival bias may lead to a lower observed incident CKD. However, the estimated annual incidence of 1.17% in our cohort was similar to the annualized rates reported by general population studies in other Asian countries.[10–13] Additionally, the supplementary analysis using IPW to account for attrition identified the same risk factors with similar risk estimates as the main analysis. The analysis of excess risk may be biased by unmeasured confounders or residual confounding of measured variables,[46] while confounder-mediator confounding is not accounted for. In addition, the PAR assumes a causal relationship between the risk factor and the outcome and the independence of the risk factors. Hence there may be concern about the validity of the formula to estimate PARs where confounding of the exposure-disease association exists.[47] Since the incidence was <10% in Chinese and Indians and 10% in Malays, the adjusted OR was used to approximate adjusted relative risk (RR) in an alternative expression which remains valid result in the presence of confounders.[47] These PARs were similar to the original estimates, while those obtained using results from the multivariable LR model[48 49] were more conservative (online supplemental etable 8). Although the PAR is an epidemiologic measure to assess the public health impact of risk factor exposure in the population, the reality is that the risk factor is unlikely to be completely eradicated. Instead, other measures such as the generalized impact fraction can estimate the fractional reduction of cases that would result from reducing the risk factor prevalence.[48 50]

In conclusion, our prospective population-based cohort study in Singapore demonstrated significant ethnic disparities in incident CKD in Asians that were partially explained by clinical, socioeconomic and behavioural factors using traditional LR and machine learning techniques. These findings may have important implications in terms of informing policy development and resource allocation in a culturally competent healthcare system to target risk factors that will bring about the greatest reduction in incident CKD.

Author affiliations
[1]Department of Renal Medicine, Singapore General Hospital, Singapore
[2]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore
[3]Department of Statistics and Applied Probability, National University of Singapore, Singapore
[4]Ophthalmology and Visual Sciences Academic Clinical Program, Duke-NUS Medical School, Singapore
[5]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Contributors CS is the guarantor and accepts full responsibility for the work and/or the conduct of the study, had access to the data, and controlled the decision to publish. CCL and CS conceptualized the study and wrote the first draft. FH performed statistical analysis. All authors interpreted the results and approved the final manuscript.

Funding This study was supported by the National Medical Research Council, NMRC/STaR/016/2013, NMRC/CIRG/1371/2013, NMRC/CIRG/1417/2015 and OFLCG/001/2017.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study (SEED) was conducted according to the Declaration of Helsinki and was approved by the Singapore Eye Research Institute Review Board and the SingHealth Centralised Institutional Review Board (2018/2717, 2018/2921, 2018/2006, 2018/2594, 2018/2570, 2015/2279, 2012/487/A).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. As the study involves human participants, the data cannot be made freely available in the manuscript, the supplemental files, or a public repository due to ethical restrictions. Nevertheless, the data are available from the Singapore Eye Research Institutional Ethics Committee for researchers who meet the criteria for access to confidential data. Interested researchers can send data access requests to the Singapore Eye Research Institute using the following email address: seri@seri.com.sg.

ORCID iDs
Cynthia Ciwei Lim http://orcid.org/0000-0003-0021-4861
Charumathi Sabanayagam http://orcid.org/0000-0002-4042-4719

REFERENCES

1 Eckardt K-U, Coresh J, Devuyst O, *et al*. Evolving importance of kidney disease: from subspecialty to global health burden. *Lancet* 2013;382:158–69.
2 Kyu HH, Abate D, Abate KH, *et al*. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018;392:1859–922.

3 Wang J, Zhang L, Tang SC-W, *et al*. Disease burden and challenges of chronic kidney disease in North and East Asia. *Kidney Int* 2018;94:22–5.
4 Roth GA, Abate D, Abate KH, *et al*. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 2018;392:1736–88.
5 Small C, Kramer HJ, Griffin KA, *et al*. Non-dialysis dependent chronic kidney disease is associated with high total and out-of-pocket healthcare expenditures. *BMC Nephrol* 2017;18:3.
6 CMMD O'Seaghdha, Lyass AP, Massaro JMP. A risk score for chronic kidney disease in the general population. *Am J Med* 2012;125:270–7.
7 Kovesdy CP, Alrifai A, Gosmanova EO, *et al*. Age and outcomes associated with BP in patients with incident CKD. *Clin J Am Soc Nephrol* 2016;11:821–31.
8 Goek O-N, Prehn C, Sekula P, *et al*. Metabolites associate with kidney function decline and incident chronic kidney disease in the general population. *Nephrol Dial Transplant* 2013;28:2131–8.
9 Rebholz CM, Coresh J, Grams ME, *et al*. Dietary acid load and incident chronic kidney disease: results from the ARIC study. *Am J Nephrol* 2015;42:427–35.
10 Lee C, Yun H-R, Joo YS, *et al*. Framingham risk score and risk of incident chronic kidney disease: a community-based prospective cohort study. *Kidney Res Clin Pract* 2019;38:49–59.
11 Chang TI, Lim H, Park CH, *et al*. Associations of systolic blood pressure with incident CKD G3-G5: a cohort study of South Korean adults. *Am J Kidney Dis* 2020;76:224–32.
12 Yano Y, Fujimoto S, Kramer H, *et al*. Long-term blood pressure variability, new-onset diabetes mellitus, and new-onset chronic kidney disease in the Japanese general population. *Hypertension* 2015;66:30–6.
13 Chien K-L, Lin H-J, Lee B-C, *et al*. A prediction model for the risk of incident chronic kidney disease. *Am J Med* 2010;123:836–46.
14 Low S, Lim SC, Zhang X, *et al*. Development and validation of a predictive model for chronic kidney disease progression in type 2 diabetes mellitus based on a 13-year study in Singapore. *Diabetes Res Clin Pract* 2017;123:49–54.
15 Sabanayagam C, Lim SC, Wong TY, *et al*. Ethnic disparities in prevalence and impact of risk factors of chronic kidney disease. *Nephrol Dial Transplant* 2010;25:2564–70.
16 Elsevier. Socioeconomic factors and racial disparities in kidney disease outcomes. *Semin Nephrol* 2013.
17 Betancourt JR, Green AR, Carrillo JE. Defining cultural competence: a practical framework for addressing racial/ethnic disparities in health and health care. *Public Health Rep* 2016.
18 Hill NR, Fatoba ST, Oke JL. Global prevalence of chronic kidney disease – a systematic review and meta-analysis. *PLoS One* 2016;11.
19 Majithia S, Tham Y-C, Chee M-L, *et al*. Cohort profile: the Singapore epidemiology of eye diseases study (SEED). *Int J Epidemiol* 2021;50:41–52.
20 Majithia S, Tham YC, Chee ML, *et al*. Singapore Chinese eye study: key findings from baseline examination and the rationale, methodology of the 6-year follow-up series. *Br J Ophthalmol* 2020;104:610–5.
21 Rosman M, Zheng Y, Wong W, *et al*. Singapore Malay eye study: rationale and methodology of 6-year follow-up study (SiMES-2). *Clin Exp Ophthalmol* 2012;40:557–68.
22 Sabanayagam C, Yip W, Gupta P, *et al*. Singapore Indian eye study-2: methodology and impact of migration on systemic and eye outcomes. *Clin Exp Ophthalmol* 2017;45:779–89.
23 Lim CC, Chee ML, Cheng C-Y, *et al*. Simplified end stage renal failure risk prediction model for the low-risk general population with chronic kidney disease. *PLoS One* 2019;14:e0212590.
24 James PA, Oparil S, Carter BL, *et al*. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the eighth joint National Committee (JNC 8). *JAMA* 2014;311:507–20.
25 American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2010;33:S62–9.
26 American Diabetes Association. Standards of medical care in diabetes--2014. *Diabetes Care* 2014;37:S14–80.
27 Levey AS, Stevens LA, Schmid CH, *et al*. A new equation to estimate glomerular filtration rate. *Ann Intern Med* 2009;150:604–12.
28 Nelson RG, Grams ME, Ballew SH, *et al*. Development of risk prediction equations for incident chronic kidney disease. *JAMA* 2019;322:2104–14.
29 Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001:1189–232.
30 Breiman L. Random forests. *Machine Learn* 2001;45:5–32.
31 Walter SD. The estimation and interpretation of attributable risk in health research. *Biometrics* 1976;32:829–49.
32 Chia YC, Ching SM. Hypertension and the development of new onset chronic kidney disease over a 10 year period: a retrospective cohort study in a primary care setting in Malaysia. *BMC Nephrol* 2012;13:173.
33 Singapore Department of Statistics. Population and population structure 2020. Available: https://www.singstat.gov.sg/find-data/search-by-theme/population/population-and-population-structure/latest-data [Accessed 1 Dec 2020].
34 Ricardo AC, Loop MS, Gonzalez F, *et al*. Incident chronic kidney disease risk among Hispanics/Latinos in the United States: the Hispanic community health Study/Study of Latinos (HCHS/SOL). *J Am Soc Nephrol* 2020;31:1315–24.
35 Saw W-Y, Tantoso E, Begum H. Establishing multiple omics baselines for three Southeast Asian populations in the Singapore integrative omics study. *Nat Commun* 2017;8:1–11.
36 Suri VR, Majid S, Chang Y-K, *et al*. Assessing the influence of health literacy on health information behaviors: a multi-domain skills-based approach. *Patient Educ Couns* 2016;99:1038–45.
37 Griva K, Yoong RKL, Nandakumar M, *et al*. Associations between health literacy and health care utilization and mortality in patients with coexisting diabetes and end-stage renal disease: a prospective cohort study. *Br J Health Psychol* 2020;25:405–27.
38 Chan GC, Teo BW, Tay JC, *et al*. Hypertension in a multi-ethnic Asian population of Singapore. *J Clin Hypertens* 2021;23:522–8.
39 Hong CY, Chia KS, Hughes K, *et al*. Ethnic differences among Chinese, Malay and Indian patients with type 2 diabetes mellitus in Singapore. *Singapore Med J* 2004;45:154–60.
40 Sheridan SL, Halpern DJ, Viera AJ, *et al*. Interventions for individuals with low health literacy: a systematic review. *J Health Commun* 2011;16:30–54.
41 Beauchamp A, Batterham RW, Dodson S, *et al*. Systematic development and implementation of interventions to optimise health literacy and access (Ophelia). *BMC Public Health* 2017;17:230.
42 Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;319:1317–8.
43 Nusinovici S, Tham YC, Chak Yan MY, *et al*. Logistic regression was as good as machine learning for predicting major chronic diseases. *J Clin Epidemiol* 2020;122:56–69.
44 van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014;14:137.
45 Garofalo C, Borrelli S, Minutolo R, *et al*. A systematic review and meta-analysis suggests obesity predicts onset of chronic kidney disease in the general population. *Kidney Int* 2017;91:1224–35.
46 Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. *Int J Epidemiol* 2013;42:1511–9.
47 Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. *Am J Public Health* 1998;88:15–19.
48 Mansournia MA, Altman DG. Population attributable fraction. *BMJ* 2018;360:k757.
49 Brady AR. Adjusted population attributable fractions from logistic regression. *Stata Technical Bulletin* 1998;7.
50 Morgenstern H, Bursic ES. A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. *J Community Health* 1982;7:292–309.