Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integration of inflammation-response gene polymorphisms and clinical covariates in a multivariable risk model for colorectal cancer in Malays

MOHD AMINUDIN MUSTAPHA

mmaminudin@unimas.my

Universiti Malaysia Sarawak

KHAIRUL FIKRI TAMRIN

Universiti Malaysia Sarawak

SITI NURFATIMAH MOHD SHAHPUDIN

Universiti Sains Malaysia

Article

Keywords: Colorectal cancer, Interleukin-8 (IL-8 – 251 T > A), Multivariable risk prediction model, Inflammation-response gene polymorphisms, Malay population

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11449621/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Chronic inflammation plays an essential role in colorectal carcinogenesis, yet genetic susceptibility studies in underrepresented Southeast Asian populations remain scarce. We constructed and evaluated a multivariable clinical-genetic risk prediction model for colorectal cancer (CRC) in an ethnically homogeneous case-control cohort of 195 histologically confirmed cases and 214 controls of Malay descent. Six functional inflammation-response gene polymorphisms (*IL-8* − 251 T > A, *TNF-α* − 308 G > A, *ICAM-1* K469E, *ICAM-1* R241G, *IL-6* − 174 G > C, and *PPAR-γ* Pro12Ala) were evaluated alongside age and sex. Multivariable logistic regression confirmed that *IL-8* − 251 T > A is strongly and independently associated with CRC risk following mutual adjustment for concurrent genetic loci and clinical covariates (adjusted OR = 2.67, 95% CI: 1.61–4.42, $p = 0.0001$). Backward stepwise elimination derived a parsimonious model retaining *IL-8* − 251 T > A (adjusted OR = 2.65, $p = 0.0001$), *TNF-α* − 308 G > A (OR = 1.45, $p = 0.075$), and age (OR = 1.11 per year, $p < 0.0001$). The model demonstrated excellent calibration (Hosmer–Lemeshow $p = 0.883$) and high discrimination (AUC = 0.825, classification accuracy = 74.8%). These findings establish *IL-8* − 251 T > A as an independent susceptibility marker in Malays and provide proof-of-concept for integrating pathway-specific genetic markers into population-calibrated CRC risk stratification.

Introduction

Colorectal cancer (CRC) remains a paramount global oncological challenge, with disease burden and etiologic factors exhibiting marked divergence across distinct ethnic populations. In Malaysia, CRC represents an escalating public health crisis within the indigenous Malay population, ranking as the leading malignancy among Malay males (accounting for 16.9% of all male cancer diagnoses) and the second most prevalent in Malay females (10.4%), surpassed only by breast carcinoma [1]. According to the Malaysia National Cancer Registry Report 2017–2021, the age-standardised incidence rate (ASR) reached 14.5 per 100,000 in Malay men and 11.3 per 100,000 in Malay women, corresponding to an estimated lifetime risk of one in 56 and one in 77 individuals, respectively [1]. Over the past fifteen years, CRC incidence among Malays has followed an uninterrupted upward trajectory rising from 11.3 per 100,000 (2007–2011) to 14.5 per 100,000 (2017–2021) in males, alongside an increase from 8.1 to 11.3 per 100,000 in females [1]. Mechanistically, chronic, non-resolving inflammation is widely recognized as an indispensable hallmark of colorectal carcinogenesis, orchestrating a pro-tumourigenic microenvironment that facilitates cellular proliferation, immune evasion, neo-angiogenesis, and extracellular matrix remodelling through complex cytokine and adhesion-molecule cascades [2, 3].

Host genetic variation within key inflammatory mediators fundamentally modulates the intensity and duration of this mucosal inflammatory response. Among these, interleukin-8 (*IL-8*), a potent CXC chemokine, mediates neutrophil recruitment and tumour-associated angiogenesis, with its promoter variant − 251 T > A (rs4073) shown to enhance transcriptional activation and modulate cancer risk [4, 5]. Tumour necrosis factor-alpha (*TNF-α*) functions as a master orchestrator of inflammatory signalling, wherein the functional − 308 G > A promoter polymorphism (rs1800629) alters baseline transcriptional

activity [6]. Concurrently, intercellular adhesion molecule-1 (*ICAM-1*) governs leukocyte transendothelial migration and has been linked to malignant progression and tumour immune evasion, with the non-synonymous K469E (rs5498) and R241G (rs1799969) coding variants implicated in cancer predisposition [7, 8]. Furthermore, interleukin-6 (*IL-6*), a pleiotropic driver of the STAT3 signalling axis in colonic epithelial survival, harbours the − 174 G > C promoter variant (rs1800795) that regulates systemic cytokine output [9, 10], while peroxisome proliferator-activated receptor gamma (*PPAR-γ*) modulates colonic mucosal homeostasis and inflammatory gene silencing, with the Pro12Ala (34 G > C, rs1801282) variant linked to altered receptor transactivation and neoplastic risk [11, 12].

Despite strong biological plausibility, evidence linking these functional single-nucleotide polymorphisms (SNPs) to CRC risk remains conflicting across international literature and remarkably scarce within Southeast Asian populations. In Malaysia, initial candidate-gene studies provided preliminary indications of genetic susceptibility: Mustapha et al. (2012) reported a greater than three-fold elevated CRC risk among homozygous *IL-8* − 251 AA carriers (OR = 3.63) [13], while Ankathil et al. demonstrated that *TNF-α* − 308 G > A conferred significant predisposition, particularly among older subjects and tobacco smokers, alongside negligible contributions from *ICAM-1* and *IL-6* variants [14, 15]. However, translating these findings has been hindered by substantial population-genetic heterogeneity. Across East and Southeast Asian populations, alleles such as *IL-6* − 174C and *PPAR-γ* 12Ala occur at drastically lower minor allele frequencies (MAFs < 5% or monomorphic) relative to European-descent cohorts, frequently yielding conflicting meta-analytic conclusions [8, 12, 16]. Crucially, prior investigations in Malaysia frequently pooled multi-ethnic samples (Malays, Chinese, and Indians) without disaggregating ethnic-specific effects or evaluating whether observed single-locus associations operate independently of concurrent inflammatory variants, leaving the distinct genetic architecture of the Malay population largely uncharacterized.

Methodologically, candidate-gene association studies have overwhelmingly relied on univariable, single-SNP testing. While informative for initial screening, this locus-by-locus approach cannot establish whether an observed association reflects a genuine independent biological effect or mutual confounding by concurrent genetic variants and clinical covariates. Furthermore, single-locus metrics provide no insight into combined discriminative ability or model calibration—benchmarks that are essential under contemporary genetic epidemiology frameworks (such as the TRIPOD guidelines) and serve as a pragmatic precursor to pathway-specific polygenic risk modelling [17–19]. Multivariable logistic regression directly addresses these limitations by providing mutually adjusted effect estimates and quantifying the cumulative discriminative performance achieved when targeted genetic variants are integrated with established clinical factors such as age and sex.

To address these gaps, this study utilizes the Malay case-control cohort from our prior investigations [13, 14] to implement a comprehensive multivariable predictive modelling framework. Specifically, our objectives were to: (i) ascertain whether the susceptibility conferred by *IL-8* − 251 T > A remains statistically robust and independent when simultaneously adjusted for other candidate inflammation SNPs (*TNF-α*, *ICAM-1*, *IL-6*), age, and sex; (ii) derive a parsimonious risk model via backward stepwise

elimination; and (iii) rigorously evaluate the resulting model's discrimination, calibration, and overall goodness-of-fit.

Results

Cohort characteristics and demographic profile

The final analytical cohort comprised 195 histopathologically confirmed sporadic colorectal cancer (CRC) cases and 214 healthy, cancer-free controls of verified Malay ethnicity (total n = 409). The demographic characteristics of the participants are summarized in Table 1. Cases were significantly older than controls (mean age 60.1 ± 12.4 vs. 45.2 ± 11.8 years, $p < 0.0001$). A slightly higher proportion of males was observed among cases (52.3%) than controls (43.5%), though this difference did not attain statistical significance ($p = 0.091$).

Table 1
Demographic characteristics of the Malay case-control sample.

Characteristic	Cases (n = 195)	Controls (n = 214)	p-value
Age, years (mean ± SD)	60.1 ± 12.4	45.2 ± 11.8	< 0.0001
Male, n (%)	102 (52.3%)	93 (43.5%)	0.091

Given the marked age difference between cases and controls, patient age was incorporated as a continuous covariate across all subsequent adjusted univariable and multivariable regression models.

Genotype distributions, Hardy–Weinberg equilibrium, and quality control

Genotype frequencies and Hardy–Weinberg equilibrium (HWE) testing in the control group across all six candidate polymorphisms are detailed in Table 2 and illustrated in Fig. 1. Genotype frequencies in the control group conformed to HWE expectations for *ICAM-1* K469E ($p = 0.972$), *ICAM-1* R241G ($p = 1.000$), *IL-6* − 174 G > C ($p = 0.570$), and *PPAR-γ* 34 C > G ($p = 1.000$). Conversely, statistically significant departures from HWE were detected in controls for *IL-8* − 251 T > A ($p < 0.0001$) and *TNF-α* − 308 G > A ($p = 0.0029$).

Table 2
Genotype counts, Hardy-Weinberg equilibrium, and allelic association for six inflammation-response gene SNPs (Malay subgroup). NC = not calculable due to zero/near-zero cell counts.

SNP	Case (Wt/Ht/V)	Control (Wt/Ht/V)	HWE p	Genotype p	Allele OR (95% CI)	Allele p
IL-8 -251 T > A	195 (26/144/25)	214 (45/161/8)	< 0.0001	0.0009	1.40 (1.06–1.85)	0.0195
TNF-α -308 G > A	195 (92/88/15)	214 (113/96/5)	0.0029	0.0363	1.32 (0.97–1.79)	0.0929
ICAM-1 K469E	195 (102/77/16)	214 (110/87/17)	0.9720	0.9707	0.98 (0.73–1.34)	0.9804
ICAM-1 R241G	195 (192/1/2)	214 (214/0/0)	1.0000	0.1905	NC	0.0574
IL-6 -174 G > C	195 (180/14/1)	214 (198/16/0)	0.5700	0.5740	1.10 (0.54–2.23)	0.9300
PPAR-γ 34 G > C	195 (192/3/0)	214 (214/0/0)	1.0000	NC	NC	NC

In the quality control evaluation, replicate bidirectional direct automated Sanger sequencing performed on a random 10% subset of samples across all genotype classes demonstrated 100% concordance with primary PCR-RFLP and AS-PCR calls, confirming high technical assay fidelity.

Univariable genetic associations and inheritance models

Univariable analyses revealed distinct association patterns across the candidate loci. The genotype distribution of *IL-8* − 251 T > A differed significantly between cases and controls (p = 0.0009), with the variant A allele conferring an elevated disease risk (allelic OR = 1.40, 95% CI: 1.06–1.85, p = 0.0195; Table 2). Similarly, *TNF-α* − 308 G > A exhibited a significant overall genotype difference (p = 0.0363), although the allelic contrast missed conventional significance (OR = 1.32, 95% CI: 0.97–1.79, p = 0.0929). In contrast, *ICAM-1* K469E and *IL-6* − 174 G > C showed no evidence of association with CRC status (p = 0.971 and p = 0.574, respectively). For *ICAM-1* R241G and *PPAR-γ* 34 C > G, homozygous variant genotypes were virtually absent across the cohort (2 and 0 carriers, respectively), precluding univariable and multivariable odds ratio calculation.

Assessment under dominant and recessive genetic models provided critical mechanistic insights into the mode of inheritance (Table 3). *IL-8* − 251 T > A demonstrated a significant dominant effect (OR = 1.73, 95% CI: 1.02–2.93, p = 0.0496) but exhibited its most pronounced susceptibility under a recessive model, wherein homozygous AA carriers possessed a nearly four-fold increased odds of CRC compared with WT/HT individuals (OR = 3.79, 95% CI: 1.67–8.61, p = 0.0009). *TNF-α* − 308 G > A followed an analogous recessive pattern: while non-significant under the dominant model (OR = 1.25, p = 0.2768), homozygous

variant (AA) carriers exhibited a more than three-fold elevated risk (OR = 3.48, 95% CI: 1.24–9.77, p = 0.0195).

Table 3
Genotype association under dominant and recessive genetic models (Malay subgroup).

SNP	Dominant OR (95% CI)	Dominant p	Recessive OR (95% CI)	Recessive p
IL-8 -251 T > A	1.73 (1.02–2.93)	0.0496	3.79 (1.67–8.61)	0.0009
TNF-α -308 G > A	1.25 (0.85–1.85)	0.2768	3.48 (1.24–9.77)	0.0195
ICAM-1 K469E	0.96 (0.65–1.42)	0.9211	1.04 (0.51–2.11)	1.0000
ICAM-1 R241G	NC	0.1075	NC	0.2267
IL-6 -174 G > C	1.03 (0.50–2.15)	1.0000	NC	0.4768
PPAR-γ 34 G > C	NC	0.1075	NC	1.0000

Adjusted univariable associations

When evaluated under a log-additive model adjusting for age and biological sex (Table 4, Fig. 2), *IL-8* −251 T > A remained highly and independently associated with CRC risk (adjusted OR = 2.59, 95% CI: 1.57–4.28, p = 0.0002). This persistence confirms that the robust association identified for *IL-8* was not an artifact of the age distribution between cases and controls. In contrast, the association for *TNF-α* −308 G > A was attenuated upon demographic adjustment (adjusted OR = 1.40, 95% CI: 0.94–2.09, p = 0.0962), while *ICAM-1* K469E and *IL-6* − 174 G > C remained entirely non-significant.

Table 4
Age- and gender-adjusted odds ratios (multivariable logistic regression, log-additive coding).

SNP	Adjusted OR*	95% CI	p-value
IL-8 -251 T > A	2.59	1.57–4.28	0.0002
TNF-α -308 G > A	1.40	0.94–2.09	0.0962
ICAM-1 K469E	1.01	0.70–1.45	0.9622
ICAM-1 R241G	NC	—	0.9982
IL-6 -174 G > C	0.98	0.43–2.23	0.9587
PPAR-γ 34 G > C	NC	—	NC

*Adjusted for age and gender.

Full multivariable model and stepwise feature selection

To resolve inter-locus confounding and evaluate mutual independence, all four polymorphic SNPs (*IL-8*, *TNF-α*, *ICAM-1*, *IL-6*), age, and sex were entered simultaneously into a full multivariable logistic regression model (n = 409). The full model exhibited significant overall fit compared with the null model (likelihood-ratio test $p < 0.0001$; Table 5, Fig. 3). In this mutually adjusted environment, *IL-8* − 251 T > A remained an independent predictor of CRC risk (adjusted OR = 2.67, 95% CI: 1.61–4.42, $p = 0.0001$). Concurrently, age remained a strong independent determinant (OR = 1.11 per additional year, 95% CI: 1.09–1.14, $p < 0.0001$). *TNF-α* − 308 G > A retained a positive association that narrowly missed statistical significance (adjusted OR = 1.45, 95% CI: 0.96–2.19, $p = 0.0802$), whereas *ICAM-1* K469E, *IL-6* − 174 G > C, and sex showed no independent contribution ($p > 0.20$).

Table 5
Full multivariable logistic regression model: mutually adjusted odds ratios for four inflammation-response gene SNPs, age, and gender (n = 409).

Variable	β (coefficient)	SE	Adjusted OR	95% CI	p-value
IL-8 -251 T > A	0.980	0.258	2.67	1.61–4.42	0.0001
TNF-α -308 G > A	0.369	0.211	1.45	0.96–2.19	0.0802
ICAM-1 K469E	0.095	0.192	1.10	0.76–1.60	0.6191
IL-6 -174 G > C	-0.195	0.433	0.82	0.35–1.92	0.6519
Age (per year)	0.105	0.011	1.11	1.09–1.14	< 0.0001
Male gender	0.304	0.242	1.36	0.84–2.18	0.2088

Backward stepwise elimination sequentially eliminated *IL-6* − 174 G > C (step 1, $p = 0.652$), *ICAM-1* K469E (step 2, $p = 0.624$), and sex (step 3, $p = 0.203$), converging on a parsimonious three-predictor model (Table 6).

Table 6
Sequence of backward stepwise elimination.

Step	Variable removed	Reason (highest p-value)	Variables remaining
1	*IL-6* -174 G > C	p = 0.652	*IL-8*, *TNF-α*, *ICAM-1*, Gender, Age
2	*ICAM-1* K469E	p = 0.624	*IL-8*, *TNF-α*, Gender, Age
3	Male gender	p = 0.203	*IL-8*, *TNF-α*, Age
4	— (stopped: all p < 0.10)	final model reached	*IL-8*, *TNF-α*, Age

In this final parsimonious model (Table 7), *IL-8* − 251 T > A maintained an effect size virtually identical to that in the full model (adjusted OR = 2.65, 95% CI: 1.61–4.38, $p = 0.0001$), alongside age (OR = 1.11 per year, $p < 0.0001$) and a suggestive contribution from *TNF-α* − 308 G > A (OR = 1.45, 95% CI: 0.96–2.18, $p = 0.0750$).

Table 7
Final parsimonious multivariable logistic regression model.

Variable	Adjusted OR	95% CI	p-value
IL-8 -251 T > A	2.65	1.61–4.38	0.0001
TNF-α -308 G > A	1.45	0.96–2.18	0.0750
Age (per year)	1.11	1.09–1.14	< 0.0001

Model performance, calibration, and incremental discriminative gain

Model performance metrics are summarized in Table 8 and Fig. 4. Both the full and parsimonious models exhibited high goodness-of-fit and satisfactory calibration, as evidenced by non-significant Hosmer–Lemeshow tests (full model: $\chi^2 = 6.35$, $p = 0.608$; final model: $\chi^2 = 3.70$, $p = 0.883$), confirming that predicted risks conformed well to observed outcomes across deciles. The parsimonious model demonstrated a more favorable balance of complexity and information retention, achieving a lower AIC (425.4 vs. 429.3) despite omitting three parameters. Discriminative accuracy remained high in both models (full model: AUC = 0.827; final parsimonious model: AUC = 0.825; Table 8, Fig. 4). At the standard 0.5 decision threshold, the model (6 covariates) achieved an overall diagnostic accuracy of 74.8% (sensitivity = 72.3%, specificity = 77.1%).

Table 8
Model performance metrics for the full and final multivariable logistic regression models.

Metric	Full model (6 covariates)	Final model (3 covariates)
Pseudo R^2 (McFadden)	0.266	0.263
AIC	429.3	425.4
Log-likelihood	−207.67	−208.70
Likelihood-ratio test p-value	< 0.0001	< 0.0001
AUC (ROC)	0.827	0.825
Hosmer-Lemeshow χ^2 (df = 8)	6.35, p = 0.608	3.70, p = 0.883
Overall classification accuracy (0.5 cutoff)	74.8%	—
Sensitivity / Specificity	72.3% / 77.1%	—

To quantify the discrete discriminative contributions of genetic variants versus clinical demographics, sensitivity modelling was conducted:

1. A genetic-only model (*IL-8* − 251 T > A and *TNF-α* − 308 G > A) yielded a baseline AUC of 0.592 (95% CI: 0.536–0.648), driven predominantly by the univariable contribution of *IL-8* (AUC = 0.573).
2. An age-only demographic model yielded an AUC of 0.808 (95% CI: 0.766–0.850).
3. The integrated clinical-genetic model (final parsimonious) achieved an AUC of 0.825 (95% CI: 0.785–0.865), delivering a statistically independent incremental gain in discrimination (ΔAUC = + 0.017).

DISCUSSION

In this study, we addressed the critical gap in genetic susceptibility research within Southeast Asian populations by constructing and evaluating a multivariable clinical-genetic risk prediction model for colorectal cancer (CRC) in an ethnically homogeneous cohort of Malays. By transitioning from conventional single-locus testing to a mutually adjusted multivariable framework, our findings establish that the IL-8 − 251 T > A polymorphism is a potent, independent genetic susceptibility marker for sporadic CRC in Malays (adjusted OR = 2.65–2.67, p = 0.0001). Furthermore, integrating targeted inflammation-pathway variants with key demographic factors yielded a parsimonious prediction model demonstrating high discriminative capacity (AUC = 0.825) and excellent calibration (Hosmer–Lemeshow p = 0.883), providing proof-of-concept for ethnically calibrated, pathway-informed risk stratification.

Our univariable and multivariable findings demonstrate that inherited susceptibility across chemokine and cytokine promoter regions in Malays is largely driven by homozygosity for transcriptional high-producer alleles. Interleukin-8 (IL-8/CXCL8), a key CXC chemokine, plays an indispensable role in colorectal tumorigenesis by promoting colonic mucosal vascular permeability, endothelial cell migration, and neutrophil infiltration, thereby fostering a pro-inflammatory and pro-angiogenic tumour microenvironment [2–5]. The functional − 251 T > A promoter polymorphism enhances transcriptional activation, and our observation of pronounced CRC risk among IL-8 − 251AA homozygotes (recessive OR = 3.79, 95% CI: 1.67–8.61, p = 0.0009) aligns with initial univariable reports by Mustapha et al. (OR = 3.63) [13] and Ankathil et al. in Malaysia [14], as well as Taiwanese cohorts reported by Tsai et al. [20], where − 251AA carriers displayed significantly elevated circulating serum IL-8 levels. While large Caucasian meta-analyses have frequently reported null associations [4, 21, 22], these discrepancies underscore substantial inter-ethnic genetic heterogeneity and highlight the importance of investigating non-European populations. Crucially, the preservation of the IL-8 effect across single-SNP, age-adjusted, and multi-locus models highlights the distinct, non-redundant role of the IL-8 chemokine axis in colorectal pathogenesis in Malays. This concurs with findings by Bondurant et al., who demonstrated in a population-based study of over 5,200 subjects that polymorphisms within the IL-8 signaling network (IL8, CXCR1, CXCR2) maintain independent associations with colorectal tumour risk after adjusting for concurrent inflammatory genes, lifestyle, and NSAID use [23], as well as multi-locus prognostic studies by Lurje et al. confirming that germline IL-8 variants retain independent predictive utility for clinical recurrence [24].

In contrast to IL-8, univariable analysis revealed an apparent recessive risk signal for TNF-α − 308AA (OR = 3.48, 95% CI: 1.24–9.77, p = 0.0195), mirroring pooled findings across Asian cohorts by Huang et al. (recessive OR = 2.23) [25] and reflecting the functional capacity of the − 308A allele to augment transcriptional output upon chronic inflammatory stimuli [6, 26]. However, upon multivariable adjustment for age, sex, and concurrent inflammatory loci, the TNF-α − 308 G > A association attenuated to an adjusted log-additive OR of 1.45 (p = 0.075–0.080). This attenuation illustrates the decisive impact of multivariable control, indicating that while TNF-α may contribute secondary susceptibility, its independent predictive contribution is modest compared to the primary driver, IL-8. Conversely, the complete lack of association for ICAM-1 K469E and IL-6 − 174 G > C is consistent with findings in Chinese and broader Malaysian cohorts [8, 14, 25], as well as solid tumour meta-analyses [7]. In particular, the minor allele frequency of IL-6 − 174C was extremely low in our cohort (3.9% overall), echoing established Asian population genomic databases where this variant is near-monomorphic [16, 17, 27]. The near-absence of variant carriers for ICAM-1 R241G and PPAR-γ 34 C > G confirms that these loci do not meaningfully contribute to colorectal cancer etiology in the Malay population.

From a predictive modelling perspective, our integrated clinical-genetic model demonstrated strong overall discrimination (AUC = 0.825) and robust calibration across risk deciles (Hosmer–Lemeshow p = 0.883). Sensitivity modelling revealed that a genetic-only model yielded a baseline AUC of 0.592, an age-only demographic model yielded an AUC of 0.808, and the integrated clinical-genetic model delivered a statistically independent incremental gain in discrimination (ΔAUC = + 0.017). This behaviour is entirely aligned with foundational principles of complex disease genetics: common, low-to-moderate penetrance susceptibility SNPs typically provide modest individual discrimination (AUC < 0.65) and are not intended to function as standalone diagnostic tests [18, 28]. Rather, their clinical utility lies in capturing an individual's inherited baseline biological vulnerability, which incrementally refines demographic and clinical risk profiles [18, 28]. The fact that IL-8 − 251 T > A maintained robust statistical significance (p = 0.0001) and added incremental discriminative value on top of age confirms that its predictive signal represents genuine biological predisposition rather than an artifact of age imbalance. In summary, combining targeted genetic markers with clinical demographics achieves high discrimination and excellent calibration, providing an initial proof-of-concept for ethnically calibrated risk stratification models to optimize colorectal cancer screening programs in Southeast Asia [1]

Two methodological aspects of our cohort warrant detailed consideration. First, colorectal cancer cases were on average 15 years older than controls (60.1 ± 12.4 vs. 45.2 ± 11.8 years, p < 0.0001). This substantial age disparity reflects the typical clinical presentation of sporadic CRC, which predominantly manifests beyond the fifth decade of life [1]. Because advanced age is an established, non-modifiable primary determinant of colorectal tumourigenesis, age was incorporated as a continuous covariate across all adjusted regressions. The stability of the IL-8 effect across unadjusted, age-adjusted, and multi-locus models confirms that this association is not confounded by age; nevertheless, residual confounding cannot be entirely excluded, and future investigations should prioritize age-matched recruitment. Second, genotype distributions for IL-8 − 251 T > A (p < 0.0001) and TNF-α − 308 G > A (p = 0.0029) deviated significantly from Hardy–Weinberg equilibrium among controls. In genetic

epidemiology, departures from HWE can arise from genotyping artifacts, non-random mating, population substructure, or selection bias. Importantly, systematic genotyping errors and preferential allele dropout were excluded by our quality-control protocol, in which 10% of samples re-analyzed via automated direct Sanger sequencing yielded 100% concordance with primary assays. Because our cohort was strictly restricted to a single self-reported ethnic group, broad inter-ethnic admixture cannot explain this pattern. Rather, these deviations are most likely attributable to subtle micro-population substructuring across northern Peninsular Malay sub-lineages (e.g., Kelantan and Kedah ancestry) [29], or hospital-based volunteer sampling dynamics, where control subjects presenting for unrelated non-malignant conditions may harbor subtle selective pressures at central inflammatory loci.

The methodological design of this study possesses several distinct strengths. First, it addresses the pronounced scarcity of genetic susceptibility research in the Malay population, providing ethnic-specific effect estimates free from confounding by inter-ethnic admixture. Second, by moving beyond univariable analyses to multivariable logistic regression and backward stepwise elimination, this study resolves inter-locus confounding and demonstrates that IL-8 operates independently of concurrent inflammatory markers. Third, our models adhered to rigorous statistical modelling standards: with 195 outcome events, the events-per-variable (EPV) ratio was 32.5 in the full model and 65.0 in the parsimonious model, well surpassing the methodological threshold of 10–20 EPV recommended to prevent overfitting and coefficient shrinkage [30]. Finally, rigorous quality control with 100% Sanger sequencing validation verified the fidelity of the primary genotyping calls.

Several limitations of this study should be recognized. First, while age was adjusted for statistically and genetic variants provided independent incremental discrimination, the 15-year age difference between cases and controls represents a potential source of residual confounding; future confirmatory studies should prioritize strictly age- and sex-matched sampling. Second, control participants were hospital-based volunteers, which may introduce selection dynamics compared to population-based registries. Third, data on unmeasured environmental, dietary, and lifestyle risk factors (e.g., tobacco smoking, red meat consumption, physical inactivity, body mass index, NSAID intake) were not available, precluding multivariable adjustment for lifestyle confounders or formal gene–environment interaction testing. Fourth, our targeted panel was restricted to six functional candidate polymorphisms across five genes, representing a modest fraction of the broader inflammatory cascade, and genome-wide polygenic architectures could not be interrogated. Finally, reported performance metrics reflect in-sample derivation, underscoring the imperative for prospective external validation in independent multi-center cohorts across Southeast Asia to establish model transportability and real-world calibration in accordance with the TRIPOD guidelines [19, 31].

Conclusion

This study establishes the *IL-8* − 251 T > A polymorphism as a potent, independent genetic susceptibility marker for sporadic colorectal cancer in the previously understudied Malay population, with the greatest risk conferred by homozygosity for the variant allele. Through a multivariable logistic regression

framework, we demonstrate that the association of *IL-8* − 251 T > A remains statistically robust following mutual adjustment for concurrent candidate inflammatory loci (*TNF-α*, *ICAM-1*, *IL-6*), biological sex, and age. Backward stepwise selection derived a highly parsimonious clinical-genetic risk model comprising *IL-8* − 251 T > A, *TNF-α* − 308 G > A, and age, which achieved excellent calibration (Hosmer–Lemeshow p = 0.883) and high discriminative performance (AUC = 0.825).

Collectively, these findings reinforce the pivotal etiologic role of chemokine-driven mucosal inflammation in colorectal tumourigenesis and provide critical proof-of-concept that integrating targeted, biologically anchored inflammatory gene variants with clinical covariates can meaningfully refine individual risk prediction. While prospective multi-centre studies with rigorous age matching and external validation remain essential to confirm model transportability, this work lays an empirical foundation for developing ethnically tailored, pathway-informed risk stratification tools to enhance early detection and targeted screening strategies in Southeast Asian populations facing an escalating colorectal cancer burden.

Methods

Study design and ethical approval

This case-control study was conducted at the Human Genome Centre, School of Medical Sciences, Universiti Sains Malaysia (USM). The study protocol was reviewed and approved by the Research Review Board and the Human Research Ethics Committee of Universiti Sains Malaysia (JEPeM Ref. No. : USMKK/PPP/JEPeM [208.4(1.4)]) and received clinical research approval from the Medical Research and Ethics Committee (MREC), Ministry of Health, Malaysia (National Medical Research Register ID: NMRR-09-351-4033). All procedures were executed in strict compliance with relevant institutional guidelines and the ethical principles outlined in the Declaration of Helsinki (2008 revision). Prior to recruitment and biological specimen collection, written informed consent was obtained from all participating individuals.

Study population and participant eligibility

Colorectal cancer cases were consecutively recruited from the surgical and oncology units of Hospital Universiti Sains Malaysia (HUSM, Kelantan) and major tertiary referral centres under the Ministry of Health Malaysia, including Hospital Sultanah Bahiyah (Alor Setar, Kedah) and Hospital Raja Perempuan Zainab II (Kota Bharu, Kelantan).

Inclusion criteria for cases comprised: (i) histopathologically confirmed primary sporadic adenocarcinoma of the colon or rectum reviewed by certified pathology laboratories; (ii) age between 30 and 70 years at initial diagnosis; (iii) self-reported and verified Malay ethnicity (up to three generations without inter-ethnic admixture); and (iv) sufficient cognitive competence to participate in structured epidemiologic interviews. Exclusion criteria for cases were: (i) clinical or histopathological evidence of familial adenomatous polyposis (FAP) or hereditary non-polyposis colorectal cancer (HNPCC/Lynch

syndrome); (ii) any personal history of prior malignancy; and (iii) documented chronic systemic inflammatory, autoimmune, or immunodeficiency disorders.

Controls were healthy, unrelated Malay volunteers recruited from the same geographic regions, comprising individuals attending outpatient clinics or blood donation drives at participating hospitals for minor, non-oncological, and non-inflammatory conditions. Controls were confirmed to have no personal or family history of colorectal cancer, gastrointestinal malignancies, or chronic autoimmune and inflammatory diseases.

A well-defined, ethnically homogeneous cohort of 195 confirmed sporadic CRC cases and 214 healthy controls of Malay descent had complete genotype and clinical covariate data and was specifically evaluated in this study.

DNA extraction and quality control

Peripheral venous blood samples (approximately 3–5 mL) were collected in sterile ethylenediaminetetraacetic acid (EDTA) tubes and stored at − 20°C until processing. Genomic DNA was isolated from peripheral blood leukocytes using the QIAamp DNA Blood Mini Kit (QIAGEN, Hilden, Germany) according to the manufacturer's protocol. DNA purity and yield were assessed spectrophotometrically (NanoDrop Technologies, Wilmington, DE, USA), with an absorbance ratio (A260/A280) between 1.8 and 2.0 considered indicative of high-grade genomic DNA.

Genotyping assays

Six functional candidate single-nucleotide polymorphisms (SNPs) across five inflammation-response genes were investigated: *IL-8* − 251 T > A (rs4073), *TNF-α* − 308 G > A (rs1800629), *ICAM-1* K469E (rs5498), *ICAM-1* R241G (rs1799969), *IL-6* − 174 G > C (rs1800795), and *PPAR-γ* Pro12Ala (34 C > G, rs1801282). Four polymorphisms (*TNF-α* − 308 G > A, *ICAM-1* K469E, *IL-6* − 174 G > C, and *PPAR-γ* 34 C > G) were genotyped using Polymerase Chain Reaction-Restriction Fragment Length Polymorphism (PCR-RFLP), followed by appropriate restriction enzyme digestion and separation on ethidium bromide-stained agarose or polyacrylamide gels. The remaining two variants (*IL-8* − 251 T > A and *ICAM-1* R241G) were determined using Allele-Specific PCR (AS-PCR) with sequence-specific primers.

To ensure rigorous quality control and genotyping accuracy, laboratory personnel were blinded to the clinical status (case vs. control) of all samples. Furthermore, a random 10% subset of samples across all genotype categories for each SNP was re-analyzed via bidirectional direct Sanger automated sequencing (First BASE Laboratories Sdn Bhd, Selangor, Malaysia). Replicate sequencing yielded a 100% concordance rate with the primary PCR-RFLP and AS-PCR calls.

Model variables and genotype coding

Candidate predictors and clinical covariates were predefined prior to model construction (Table 9). Four SNPs—*IL-8* − 251 T > A, *TNF-α* − 308 G > A, *ICAM-1* K469E, and *IL-6* − 174 G > C—were coded as ordinal predictors under a log-additive (per-allele) genetic model: wild-type homozygote (Wt) = 0, heterozygote (Ht) = 1, and variant homozygote (V) = 2. Two SNPs (*ICAM-1* R241G and *PPAR-γ* 34 C > G) exhibited near-monomorphic distributions with fewer than three variant-allele carriers across the entire cohort; to prevent severe numerical instability arising from complete or quasi-complete data separation, these two loci were excluded from multivariable model building and reported separately via univariable exact tests. Baseline demographic covariates included age at diagnosis/enrolment (continuous, years) and biological sex (binary: female = 0, male = 1). The dependent primary outcome was binary disease status (case = 1, control = 0).

Table 9
Variables and coding used in the multivariable logistic regression model.

Variable	Coding	Role in model
CRC status (outcome)	0 = Control, 1 = Case	Dependent variable
IL-8 -251 T > A	Wt = 0, Ht = 1, V = 2 (log-additive)	Predictor
TNF-α -308 G > A	Wt = 0, Ht = 1, V = 2 (log-additive)	Predictor
ICAM-1 K469E	Wt = 0, Ht = 1, V = 2 (log-additive)	Predictor
IL-6 -174 G > C	Wt = 0, Ht = 1, V = 2 (log-additive)	Predictor
Age	Continuous, years	Covariate
Gender	0 = Female, 1 = Male	Covariate

Statistical analysis and model development

Analyses were performed in Python 3.12 using SciPy 1.17, statsmodels 0.15, and scikit-learn 1.8. In the logistic regression manuscript, statsmodels was used for the model fitting and Hosmer-Lemeshow was coded manually. scikit-learn was used only for the ROC and AUC calculations and the classification table. SciPy handled the chi-square and Fisher's exact tests, and statsmodels handled the adjusted regression.

i. Univariable genetic association

Hardy-Weinberg equilibrium (HWE) in the control group was assessed for each SNP using the chi-square goodness-of-fit test. Genotype and allele distributions between cases and controls were compared using Pearson's chi-square test or Fisher's exact test when expected cell counts were < 5. Crude odds ratios (ORs) and 95% confidence intervals (CIs) were estimated for allelic contrasts as well as dominant (Ht + V vs. Wt) and recessive (V vs. Wt + Ht) inheritance models. Two-tailed p-values < 0.05 were deemed

nominally significant, and a Bonferroni-corrected significance threshold of $p < 0.0083$ (0.05 / 6 SNPs) was applied to adjust for multiple testing.

ii. Multivariable modelling and variable selection

A full multivariable logistic regression model was initially fitted incorporating all four informative SNPs, age, and sex as concurrent independent predictors. Backward stepwise elimination was then applied to derive a parsimonious prediction model. At each iteration, the covariate with the highest Wald p-value was removed sequentially until all retained variables exhibited $p < 0.10$. Patient age was retained a priori in all intermediate and final models regardless of p-value, given its established role as an independent demographic determinant of colorectal carcinogenesis and to adjust for baseline age differences between groups.

iii. Model performance, calibration, and discrimination:

In accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [19], model performance was comprehensively evaluated across multiple statistical dimensions. Overall goodness-of-fit and relative model quality were quantified using McFadden's pseudo R^2, model log-likelihood, likelihood-ratio tests (LRT), and the Akaike Information Criterion (AIC). Model calibration across risk deciles was assessed using the Hosmer–Lemeshow goodness-of-fit test (10 groups), where $p > 0.05$ indicated satisfactory agreement between observed outcomes and predicted risks. Discriminative capacity was determined by the area under the receiver operating characteristic curve (ROC-AUC) with 95% confidence intervals, alongside classification accuracy, sensitivity, and specificity computed at a default decision threshold of 0.5. Furthermore, the risk of model overfitting and optimism was evaluated using the events-per-variable (EPV) ratio, benchmarked against the standard methodological threshold of ≥ 10–20 EPV [28].

Declarations

Conflicts of Interest

The authors declare that they have no known competing financial or non-financial interests that could influence the work reported in this paper, either directly or indirectly.

Declaration of Generative AI And AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the authors used Gemini and Claude AI in order to assist with language editing, refinement of phrasing, and structuring of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

This work was financially supported by Universiti Sains Malaysia under the Research University Grant scheme (Grant No. 1001/PPSP/812032).

Author Contribution

M.A.M: Data curation, Methodology, Formal analysis, Writing – original draft, Writing – review and editing. K.F.T: Conceptualization, Methodology, Formal analysis, Writing – review and editing. S.N.M.S: Data curation, Methodology, Formal analysis, Writing – review and editing.

Acknowledgement

We dedicate this work to the memory of the late Dr. Ravindran Ankathil, who initiated, led, and supervised this research project with unwavering guidance. The authors wish to express their heartfelt thanks to Dr. B. M. Biswal and Dr. V. M. K. Bhavaraju (Hospital Universiti Sains Malaysia), and Dr. Ahmad Shanwani and Dr. Zaidi Zakaria (Hospital Raja Perempuan Zainab II, Kota Bharu) for their essential support in clinical coordination and patient recruitment. We are also deeply grateful to all participants and hospital staff who contributed to this study.

Data Availability

The datasets generated and/or analysed during the current study are available in the European Nucleotide Archive (ENA) repository under study accession PRJEB129248 (ERP208583), titled “Integration of Inflammation-Response Gene Variants and Clinical Factors in a Multivariable Colorectal Cancer Risk Model for Malays”, released on 30 October 2026. The dataset can be accessed through the accession number PRJEB129248 (ERP208583).

References

1. Yusof, S.N., *Malaysia National Cancer Registry Report (MNCR) 2017-2021*, M.Z.M.K. Waramlah Ramli, Adibah Ibrahim, Noor Rafizah Aminah Aziz, Izwana Hamzah, Tan Hooi Shyuan, Dr. Siti Mariam Ja'afar, Sukumaran Raman, Dr. Umawathy Sundrarajoo, Nurulaisyah Hamzah, Balqis Bahtiar, Editor. 2025, National Cancer Registry Department, Institut Kanser Negara: Putrajaya. p. 90.
2. Huang, H.-W., et al., *Association between inflammation and function of cell adhesion molecules influence on gastrointestinal cancer development.* Cells, 2021. **10**(1): p. 67.
3. Schmitt, M. and F.R. Greten, *The inflammatory pathogenesis of colorectal cancer.* Nature Reviews Immunology, 2021. **21**(10): p. 653-667.
4. Hu, L.X., et al., *Lack of association between interleukin-8-251 T>A polymorphism and colorectal cancer risk: a meta-analysis based on 3,019 cases and 3,984 controls.* Asian Pac J Cancer Prev, 2012. **13**(10): p. 5075-9.

5. Wang, N., et al., *-251 T/A polymorphism of the interleukin-8 gene and cancer risk: a HuGE review and meta-analysis based on 42 case-control studies.* Mol Biol Rep, 2012. **39**(3): p. 2831-41.
6. Balkwill, F., *Tumour necrosis factor and cancer.* Nat Rev Cancer, 2009. **9**(5): p. 361-71.
7. Cheng, D. and B. Liang, *Intercellular Adhesion Molecule-1 (ICAM-1) Polymorphisms and Cancer Risk: A Meta-Analysis.* Iran J Public Health, 2015. **44**(5): p. 615-24.
8. Qiu, Z., et al., *Roles of intercellular cell adhesion molecule-1 (ICAM-1) in colorectal cancer: expression, functions, prognosis, tumorigenesis, polymorphisms and therapeutic implications.* Front Oncol, 2022. **12**: p. 1052672.
9. Hirano, T., *IL-6 in inflammation, autoimmunity and cancer.* Int Immunol, 2021. **33**(3): p. 127-148.
10. Terry, C.F., V. Loukaci, and F.R. Green, *Cooperative influence of genetic polymorphisms on interleukin 6 transcriptional regulation.* J Biol Chem, 2000. **275**(24): p. 18138-44.
11. Deeb, S.S., et al., *A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity.* Nat Genet, 1998. **20**(3): p. 284-7.
12. Wei, Z., G. Han, and X. Bai, *Effect of Proliferator-Activated Receptor-γ Pro12Ala Polymorphism on Colorectal Cancer Risk: A Meta-Analysis.* Med Sci Monit, 2015. **21**: p. 1611-6.
13. Mustapha, M.A., et al., *Risk modification of colorectal cancer susceptibility by interleukin-8-251T> A polymorphism in Malaysians.* World Journal of Gastroenterology: WJG, 2012. **18**(21): p. 2668.
14. Ankathil, R., et al., *Contribution of Genetic Polymorphisms of Inflammation Response Genes on Sporadic Colorectal Cancer Predisposition Risk in Malaysian Patients - A Case Control Study.* Asian Pac J Cancer Prev, 2019. **20**(6): p. 1621-1632.
15. Lim, K.G., *A review of colorectal cancer research in malaysia.* Med J Malaysia, 2014. **69 Suppl A**: p. 23-32.
16. Wu, D., et al., *Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore.* Cell, 2019. **179**(3): p. 736-749.e15.
17. Chatterjee, N., J. Shi, and M. García-Closas, *Developing and evaluating polygenic risk prediction models for stratified disease prevention.* Nat Rev Genet, 2016. **17**(7): p. 392-406.
18. Collins, G.S., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement.* Ann Intern Med, 2015. **162**(1): p. 55-63.
19. Steyerberg, E.W., et al., *Assessing the performance of prediction models: a framework for traditional and novel measures.* Epidemiology, 2010. **21**(1): p. 128-38.
20. Tsilidis, K.K., et al., *Association of common polymorphisms in IL10, and in other genes related to inflammatory response and obesity with colorectal cancer.* Cancer Causes Control, 2009. **20**(9): p. 1739-51.
21. Min, L., et al., *Tumor necrosis factor-a polymorphisms and colorectal cancer risk: a meta-analysis.* PLoS One, 2014. **9**(1): p. e85187.
22. Walczak, A., et al., *The IL-8 and IL-13 gene polymorphisms in inflammatory bowel disease and colorectal cancer.* DNA Cell Biol, 2012. **31**(8): p. 1431-8.

23. Bondurant, K.L., et al., *Interleukin genes and associations with colon and rectal cancer risk and overall survival.* Int J Cancer, 2013. **132**(4): p. 905-15.
24. Lurje, G., et al., *Polymorphisms in VEGF and IL-8 predict tumor recurrence in stage III colon cancer.* Ann Oncol, 2008. **19**(10): p. 1734-41.
25. Jamal, R., et al., *Cohort Profile: The Malaysian Cohort (TMC) project: a prospective study of non-communicable diseases in a multi-ethnic population.* Int J Epidemiol, 2015. **44**(2): p. 423-31.
26. Qiu, Z., et al., *Evaluation of ICAM-1 rs5498 and rs3093030 Polymorphisms in Chinese Patients with Colorectal Cancer.* DNA Cell Biol, 2021. **40**(2): p. 384-392.
27. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.
28. Janssens, A.C. and C.M. van Duijn, *Genome-based prediction of common diseases: advances and prospects.* Hum Mol Genet, 2008. **17**(R2): p. R166-73.
29. Tsai, C.W., et al., *The Significant Impacts of Interleukin-8 Genotypes on the Risk of Colorectal Cancer in Taiwan.* Cancers (Basel), 2023. **15**(20).
30. Hatin, W.I., et al., *Population genetic structure of peninsular Malaysia Malay sub-ethnic groups.* PLoS One, 2011. **6**(4): p. e18312.
31. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat Med, 1996. **15**(4): p. 361-87.

Figures

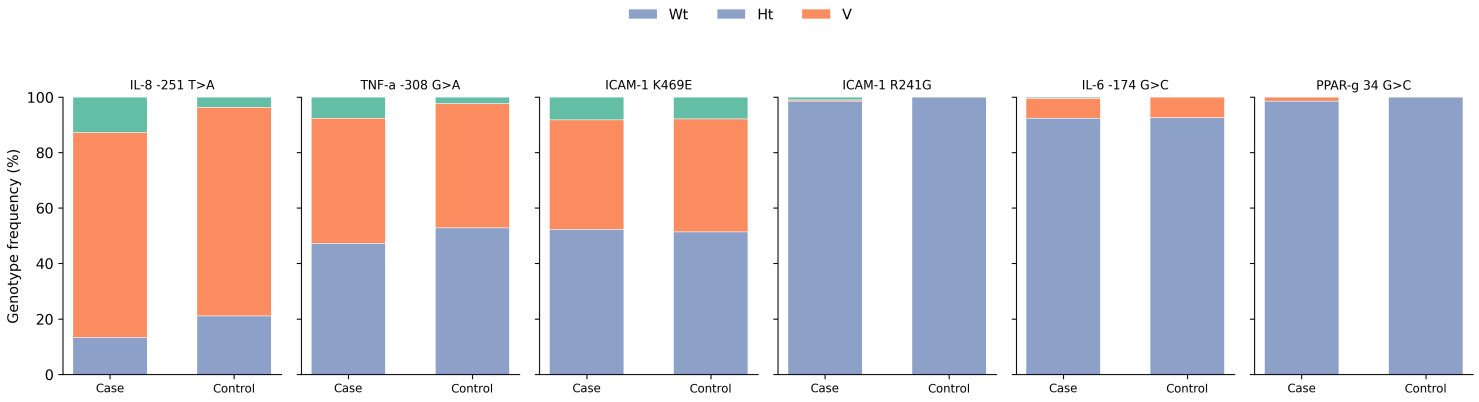


Figure 1

Stacked bar chart of genotype distribution (Wt/Ht/V, %) for six inflammation-response gene SNPs in CRC cases versus controls, Malay subgroup.

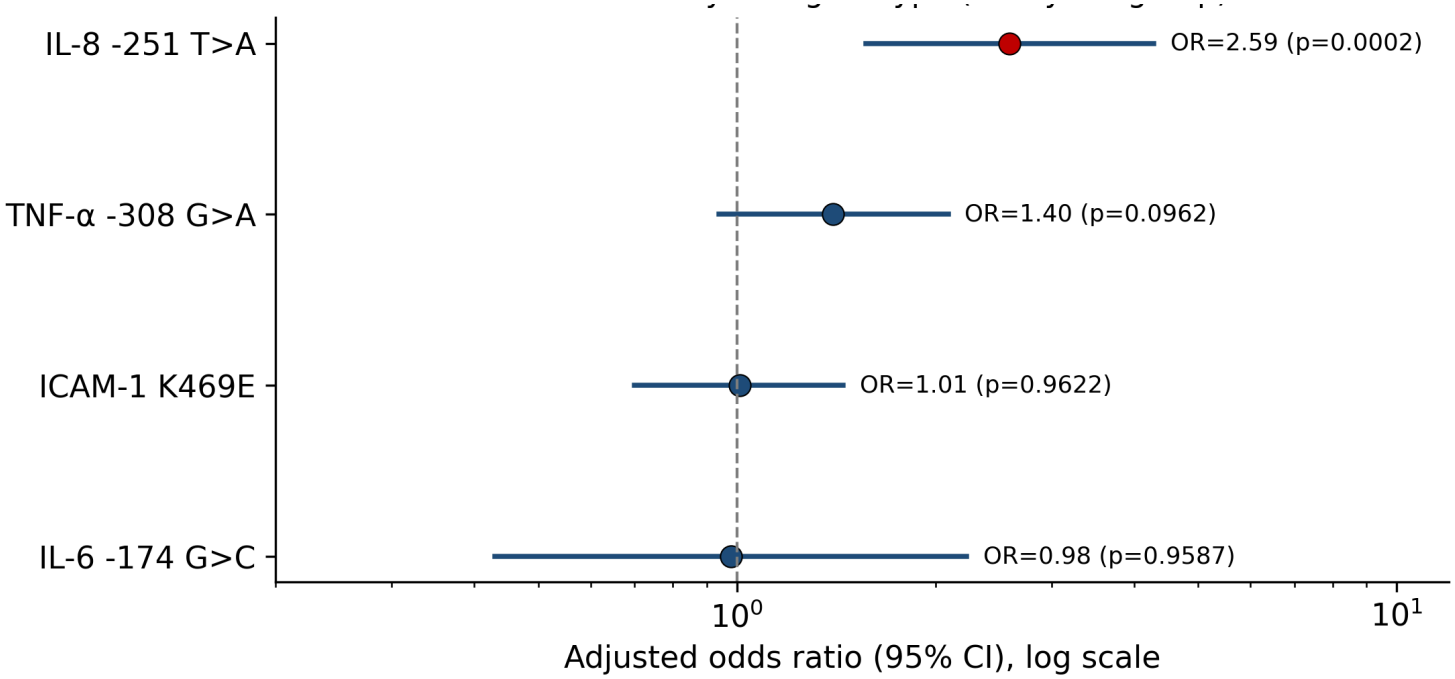


Figure 2

Forest plot of age- and gender-adjusted odds ratios (95% CI, log scale) for CRC risk by SNP genotype, Malay subgroup. *ICAM-1* R241G and *PPAR-γ* 34 G>C are omitted owing to unstable estimates from sparse variant counts (see Table 3).

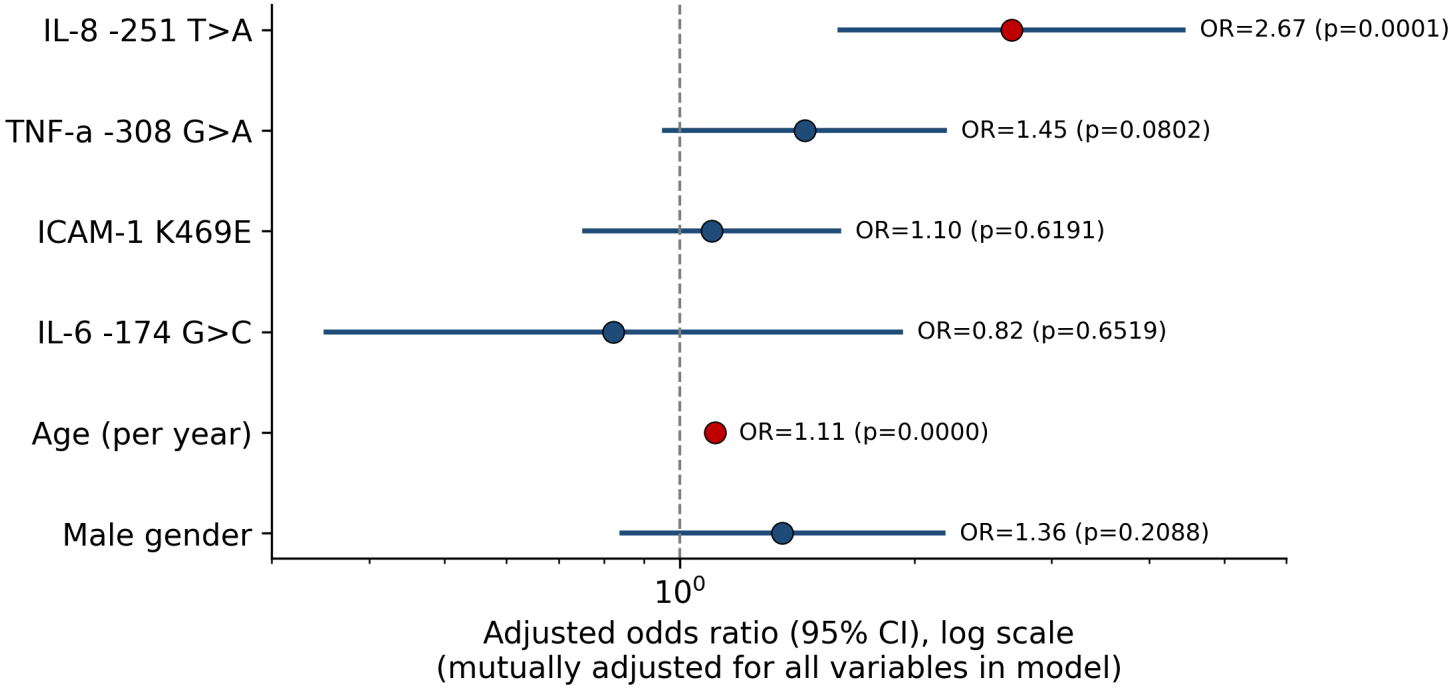


Figure 3

Forest plot of the full multivariable logistic regression model (mutually adjusted odds ratios with 95% CI, log scale), Malay subgroup. Red markers indicate p<0.05.

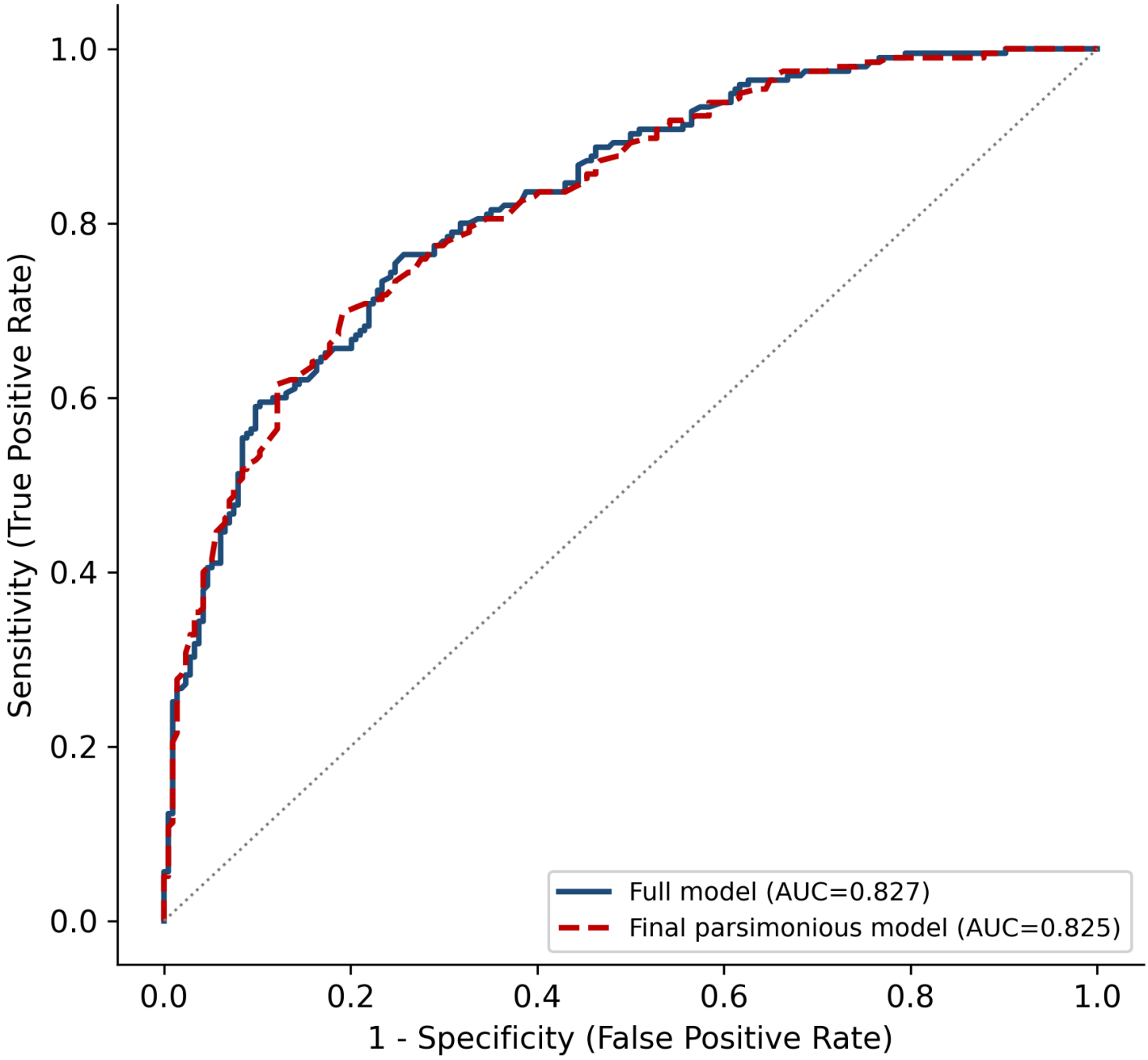


Figure 4

Receiver operating characteristic (ROC) curves for the full multivariable model (four SNPs + age + gender) and the final parsimonious model (*IL-8* -251 T>A + *TNF-α* -308 G>A + age), Malay subgroup.