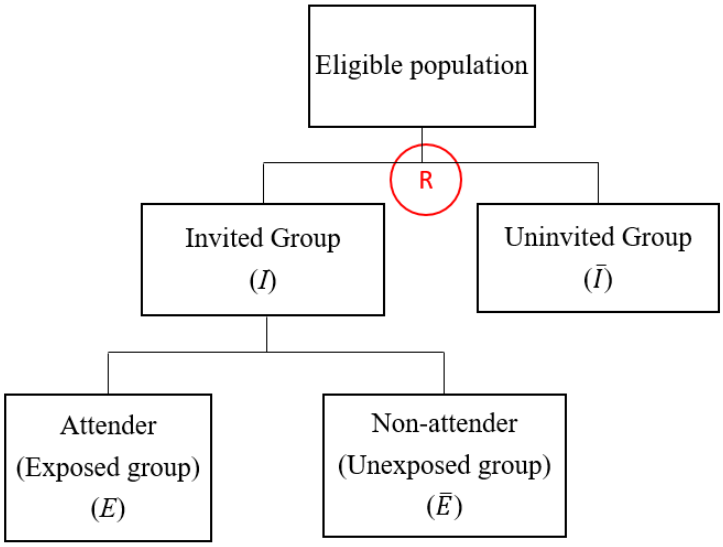
Supplementary Materials: Self-selection bias adjustment and Bayesian Poisson regression model in population-based cancer service screening

Population-based cancer service screening is often faced with self-selection bias. Namely, those who had participated in the screening are different from those who had not in terms of basic characteristics such as socio-economic status. To adjust such a kind of self-selection bias, it is imperative to follow intention-to-treat (ITT) analysis under the framework of a randomized controlled trial (RCT) design. We herein begin with the illustration of ITT analysis with a hypothetical example of population-based fecal immunochemical test (FIT) screening with a RCT design as shown in the following Supplementary Figure 1.



Supplementary Figure 1. The conceptual diagram showing study design for evaluating the effectiveness of a population-based FIT screening with a RCT design

Relative risk of measuring effectiveness with ITT analysis in population-based FIT screening

In order to assess whether colorectal cancer (CRC) screening with FIT can reduce the incidence of advanced-stage CRC or CRC mortality, the eligible population are randomized into two groups, the invited group (receive screen with FIT) and the uninvited group (without FIT), denoted by I (the invited group) and $\bar{I}$ (the uninvited group) in the Supplementary Figure 1, respectively. We here use CRC death as an example but the similar derivation can be also applied to the advanced-stage CRC. Even in the RCT design, the invited group is also classified into two groups, the exposed group (attender) and the unexposed group (non-attender) according to attendance rate. The relative risk of CRC mortality between group I and group $\bar{I}$ is expressed as:

$$\frac{P(CRC\ death\ in\ the\ invited\ group\ I)}{P(CRC\ death\ in\ the\ uninvited\ group\ \bar{I})} = \frac{P(D|I)}{P(D|\bar{I})} \qquad \text{(S-1)}$$

, which is used to unbiasedly measure the reduction of death from CRC in light of ITT analysis to reduce self-selection bias in light of per protocol analysis if the corresponding relative risk between the exposed group and the uninvited group is used because those who are willing to attend the screen would not be representative of the control group derived from the underlying population. Note that relative risk is changed to relative rate when person-years is used for the denominator.

Chiu H-M, *et al. Gut* 2021; 70:2321–2329. doi: 10.1136/gutjnl-2020-322545

Relative rate of measuring effectiveness with ITT analysis in population-based FIT service screening

Unlike the RCT design, there is lacking of the uninvited group (the control group $\bar{I}$ in Supplementary Figure 1) in the evaluation of population-based FIT service screening as all the eligible population are invited for FIT screening. The comparator used for evaluation often relies on the unexposed group (see Supplementary Figure 1). It is more prone to self-selection bias when one would like to use the relative rate of being dead from CRC between the exposed group (E) and the unexposed group ($\bar{E}$) to measure the effectiveness of FIT screening with the following expression: $\frac{P(CRC\ death\ in\ the\ exposed\ group\ E)}{P(CRC\ death\ in\ the\ unexposed\ group\ \bar{E})}$

To adjust for such a kind of self-selection bias, one has to return to use the ITT analysis to get unbiased estimated effectiveness in population-based FIT service screening as if obtained from a RCT design by making use of screening characteristics from uptake screening, referral rate for diagnostic examination, until the cecal intubation rate with complete colonoscopy.

We begin with self-selection in relation to the uptake of screen (screening arte) alone. The numerator regarding the risk for being dead from CRC in the invited group of (S-1) can be decomposed into two parts, the exposed group and the unexposed

group with the mathematical formula expressed as:

$$\mathrm{RR} = \frac{P(CRC\ death\ in\ the\ invited\ group\ I)}{P(CRC\ death\ in\ the\ uninvited\ group\ \bar{I})} = \frac{P(CRC\ death|Invited)}{P(CRC\ death|Uninvited)}$$

$$= \frac{P(D|I)}{P(D|\bar{I})} = \frac{P(D|E,I)P(E|I)+P(D|\bar{E},I)P(\bar{E}|I)}{P(D|\bar{I})}$$

where D, E, $\bar{E}$, I and $\bar{I}$ represents dead from CRC, exposed, not exposed, invited, and not invited to FIT screening, respectively.

Since we know the status of exposure to screen, the invitation (I) conveys little information, which is so-called conditional independence that simplifies $P(D|E,I)$ and $P(D|\bar{E},I)$ into $P(D|E)$ and $P(D|\bar{E})$. Consider $P(E|I)$ and $P(\bar{E}|I)$ as the screening rate (r_E) and the complementary of screening rate (1-r_E)

The equation above is reduced to

$$\frac{P(D|E)P(E|I)+P(D|\bar{E})P(\bar{E}|I)}{P(D|\bar{I})} = \frac{P(D|E)}{P(D|\bar{I})} \times r_E + \frac{P(D|\bar{E})}{P(D|\bar{I})} \times (1 - r_E) \qquad \text{(S-2)}$$

Let

$RR_E \left(= \frac{P(CRC\ death\ in\ the\ exposed\ group\ E)}{P(CRC\ death\ in\ the\ uninvited\ group\ \bar{I})} = \frac{P(D|E)}{P(D|\bar{I})}\right)$ and

$RR_{\bar{E}} \left(= \frac{P(CRC\ death\ in\ the\ exposed\ group\ E)}{P(CRC\ death\ in\ the\ uninvited\ group\ \bar{I})} = \frac{P(D|\bar{E})}{P(D|\bar{I})}\right)$ be defined as two relative rates of CRC mortality in the exposed and unexposed groups compared to the control group, respectively. Given ITT analysis, the final part of the equation (S-2) gives the

estimate of first adjusted RR (aRR_1) taken as the average of RR_E and $RR_{\bar{E}}$ weighted by the screening rate.

$$\text{aRR}_1 \cong RR_E \times r_E + RR_{\bar{E}} \times (1 - r_E) \quad \text{(S-3)}$$

Note that $P(D|\bar{I})$ represents the expected mortality rate of CRC in the absence of screening, equivalent to the mortality rate of CRC of the uninvited (control) group in the randomized controlled trial. In population-based FIT service screening, it is impossible to have the uninvited group as seen in the RCT, the pre-screening group with the adjustment for annual natural growth rate of increasing incidence rate of CRC as mentioned in the text of the statistical part in methods section was used for a proxy for the risk of being dead from CRC in the invited group.

Relative rate of advanced CRC or CRC death with the full adjustment from uptake screening until the completeness of colonoscopy

Figure 1 in the main text shows a cascade of processes from uptake screening, the referral for confirmatory diagnostic examination until complete colonoscopy in population-based FIT service screening given the population is invited. The invited group in the equation (S-1) was further classified into six groups.

Therefore, the numerator of equation (S-1) on the risk for being dead from CRC in the invited group is first divided into two parts as in the equation (S-2), the exposed

group and the unexposed group. The exposed group part is further decomposed into five corresponding parts. The expression is written as:

$$P(D|I) = P(D, E, Po, Re, Cp, Ci|I) + P(D, E, Po, Re, Cp, \overline{Ci}|I) + P(D, E, Po, Re, \overline{Cp}|I) + P(D, E, Po, \overline{Re}|I) + P(D, E, \overline{Po}|I) + P(D, \bar{E}|I) \quad \text{(S-4)}$$

The first five parts on the right side of equation are the decomposition of the conditional probability ($P(D, E|I)$) for those who exposed to screening (E) after invitation (I) in opposite to the counterpart ($P(D, \bar{E}|I)$) for the unexposed group ($\bar{E}$) after invitation (I). Note that as whether to have complete colonoscopy to reach the cecum can only be assessed conditioned on the fact that positive-test subjects were referred to colonoscopy after they had the uptake of FIT screening, all these conditional probabilities are further expanded in a forward manner following a cascade of subsequent dichotomous outcomes for positive FIT test (Po), referral (Re), colonoscopy (Cp), and complete colonoscopy (Ci). This means that the conditional probability is further expanded given the positive result of FIT. Once the negative outcome (denoted by $\overline{Po}$) is found, the expansion of the conditional probability ends.

Furthermore, as all these conditional probabilities refer to a cascade of the abovementioned characteristics and its associated death from CRC given those who were invited to screen they can be simplified by memoryless property for the outcome

earlier. Take $P(D, E, Po, Re, Cp, Ci|I)$) as an example, it can be re-expressed by the following conditional probabilities.

$$P(D, E, Po, Re, Cp, Ci|I)$$
$$= P(D|E, Po, Re, Cp, Ci, I) \times P(Ci|E, Po, Re, Cp, I) \times P(Cp|E, Po, Re, I) \times$$
$$P(Re|E, Po, I) \times P(Po|E, I) \times P(E|I) \quad \text{(S-5)}$$

Death from CRC given complete colonoscopy or not is therefore independent of events earlier. This means once information on whether to reach cecum with colonoscopy is exactly known the previous outcomes on positive FIT test, the referral to have confirmatory diagnosis, the administration of colonoscopy, cannot provide additional information. Namely, $P(D|E, Po, Re, Cp, Ci, I) = P(D|Ci)$. Recall that such a property is called conditional independence. The same idea is also applied to other parts. Equation (S-5) can then be rewritten as

$$= P(D|Ci) \times P(Ci|Cp) \times P(Cp|Re) \times P(Re|Po) \times P(Po|E) \times P(E|I) \quad \text{(S-6)}$$

Let r_{CI}, r_{CP}, r_{REF}, r_{POS}, and r_E denote cecal intubation rate (CIR), the proportion of selecting colonoscopy, referral rate to diagnostic examination, positive rate, and screening rate. Equation (S-6) can be further expressed as

$$= P(D|Ci) \cdot r_{CI} \cdot r_{CP} \cdot r_{REF} \cdot r_{POS} \cdot r_E \quad \text{(S-6)}$$

Equation (S-4) can be expressed as

$$P(D|I) = P(D|Ci) \cdot r_{CI} \cdot r_{CP} \cdot r_{REF} \cdot r_{POS} \cdot r_E +$$
$$P(D|\overline{Ci}) \cdot (1 - r_{CI}) \cdot r_{CP} \cdot r_{REF} \cdot r_{POS} \cdot r_E +$$
$$P(D|\overline{Cp}) \cdot \left(1 - r_{Cp}\right) \cdot r_{REF} \cdot r_{POS} \cdot r_E +$$
$$P(D|\overline{Re}) \cdot (1 - r_{REF}) \cdot r_{POS} \cdot r_E +$$
$$P(D|\overline{Po}) \cdot (1 - r_{POS}) \cdot r_E + P(D|\bar{E}) \cdot (1 - r_E) \quad \text{(S-7)}$$

To render the equation of (S-7) adapted for the regression model and amenable to estimation of parameters, taking into account age, gender, and increasing incidence trend, we applied the Bayesian DAG Poisson regression model proposed by Wu et al[1] for evaluating the effectiveness of Taiwanese Nationwide CRC Screening Program. The Bayesian DAG Poisson regression model links the relationship of the status of exposure to screen (1=exposed for attenders, 0=unexposed for non-attenders) with the number of advanced-stage CRC or CRC death in each group that is assumed to follow Poisson distribution. The self-selection bias using adjusted RR (aRR_1) in light of ITT analysis can be adjusted by the following Poisson regression model expressed by

$$\log(\mu) = log(PY) + \beta_b + \sum_{i=1}^{8} \beta_i X_i \quad \text{(S-8)}$$

where μ denotes the expected number of advanced-stage CRC or CRC death, PY is

the corresponding person years, $X_1 – X_6$ are indicator variables for the six groups in the invited group, X_7 for age, and X_8 for sex. Note that β_b accounts for the natural growth rate of incidence trend of CRC during the screening period had screening not taken place and was estimated as 0.0443 (se=0.000243) based on the extrapolation with time trend before the screening period.

Taking the exponent of six regression coefficients (β_1 - β_6), the corresponding regression coefficients, gives six RRs for the detailed groups compared to the uninvited group.

The self-selection bias made by using adjusted RR (aRR_2) in light of ITT analysis, making allowance for positive FIT test, referral, the choice of colonoscopy, and complete colonoscopy, was formulated as calculated as

$$aRR_2 = e^{\beta_1} \cdot r_{CI} \cdot r_{CP} \cdot r_{REF} \cdot r_{POS} \cdot r_E + e^{\beta_2} \cdot (1 - r_{CI}) \cdot r_{CP} \cdot r_{REF} \cdot r_{POS} \cdot r_E + e^{\beta_3} \cdot \left(1 - r_{Cp}\right) \cdot r_{REF} \cdot r_{POS} \cdot r_E + e^{\beta_4} \cdot (1 - r_{REF}) \cdot r_{POS} \cdot r_E + e^{\beta_5} \cdot (1 - r_{POS}) \cdot r_E + e^{\beta_6} \cdot (1 - r_E) \quad \text{(S-9)}$$

Sensitivity Analysis

In addition to estimating the adjusted relative rate as above, the equation (S-9) enables us to further calculate the relative rate of advanced-stage CRC or CRC death for the conditions:

- if all subjects with positive FIT results complying to diagnostic examination (100% referral rate, $r_{REF} = 1$) (no group 4),

$$aRR_{REF} = e^{\beta_1} \cdot r_{CI} \cdot r_{CP} \cdot r_{POS} \cdot r_E + e^{\beta_2} \cdot (1 - r_{CI}) \cdot r_{CP} \cdot r_{POS} \cdot r_E + e^{\beta_3} \cdot (1 - r_{Cp}) \cdot r_{POS} \cdot r_E + e^{\beta_5} \cdot (1 - r_{POS}) \cdot r_E + e^{\beta_6} \cdot (1 - r_E)$$

- if all subjects with positive FIT results complying to diagnostic examination with colonoscopy (100% choice of colonoscopy, ($r_{REF} = 1$ and $r_{CP} = 1$)) (without groups 3 and 4),

$$aRR_{REF} = e^{\beta_1} \cdot r_{CI} \cdot r_{POS} \cdot r_E + e^{\beta_2} \cdot (1 - r_{CI}) \cdot r_{POS} \cdot r_E + e^{\beta_5} \cdot (1 - r_{POS}) \cdot r_E + e^{\beta_6} \cdot (1 - r_A)$$

- and if all subjects with positive FIT results complying to diagnostic examination with complete colonoscopy till cecum (100% complete colonoscopy, ($r_{REF} = 1$, $r_{CP} = 1$ and $r_{CI} = 1$) (without groups 2, 3 and 4),

$$aRR_{REF} = e^{\beta_1} \cdot r_{POS} \cdot r_E + e^{\beta_5} \cdot (1 - r_{POS}) \cdot r_E + e^{\beta_6} \cdot (1 - r_E).$$

Bayesian Monte Carlo Markov Chain (MCMC) method was implemented to estimate the adjusted RRs as indicated above for the comparisons of outcome in both the exposed and the unexposed group with the comparator using the pre-screening epoch between 1998 and 2003, the period before nationwide CRC screening was launched, taking into account age, gender and the growth rate of incidence trend in the

absence of screening. Supplementary Tables 1 and 2 show the data layout used for the Poisson regression model.

In the framework of Bayesian DAG Poisson regression model, we assigned the screening rate (r_E) following a Beta distribution, $Beta(3067853, 2349846)$, where the former and latter numbers represent numbers of the exposed and the unexposed group in this study. β_b follows normal distribution, $N(0.0443, 0.000243)$. We used the non-informative priors for $\beta_1 - \beta_4$, which follow normal distribution, $N(0, 10^6)$.

For the adjustment of stage shifting, because cancer stage information was insufficient before 2003, we made us of the stage information from a study conducted by Ju et al. at Taipei Veterans General Hospital and dataset from National Taiwan University Hospital during the period of 1991 and 2000 to derive the proportion of advanced-stage CRC (AJCC stage II and higher) as 86% with the Bayesian conjugate approach.[2] (Supplementary Table 3) Accordingly, the estimated number of advanced-stage CRC was 2,387. The information on anatomical site of CRC was derived in the same way and the proportion of distal cancer was 78%. (Supplementary Table 4) Information on CRC death from distal cancer was also derived in the same manner and specified in Supplementary Table 5.

Chiu H-M, *et al. Gut* 2021; 70:2321–2329. doi: 10.1136/gutjnl-2020-322545

Supplementary Table 1

Tabular data for CRC death by the exposure to screen and the control group cross-tabulated by sex and age groups

	Number of death			**Person years**		
	Exposed	Unexposed	Pre-screened epoch (control)	Exposed	Unexposed	Pre-screened epoch (control)
Male						
50-54	45	680	116	629,192	5,822,818	706,395
55-59	259	1,710	105	1,875,536	5,417,892	418,808
60-64	401	1,879	161	1,685,326	3,430,697	391,215
65-69	488	2,045	264	1,305,304	2,451,409	323,141
70+	514	3,192	775	863,255	2,253,899	552,550
Female						
50-54	44	513	99	1,081,693	5,347,788	703,688
55-59	299	1,092	97	2,775,799	4,691,169	425,196
60-64	316	1,091	127	2,271,398	3,119,688	414,159
65-69	337	1,144	181	1,632,201	2,510,613	348,205
70+	374	2,204	462	1,059,747	2,612,399	470,792
Total	3,077	15,550	2,387	15,179,451	37,658,372	4,754,149

Supplementary Table 2

Tabular data for CRC advanced CRC by the exposure to screen and the control group

cross-tabulated by sex and age groups

	Number of advanced CRC			**Person years**		
	Exposed	Unexposed	Pre-screened epoch (control)	Exposed	Unexposed	Pre-screened epoch (control)
Male						
50-54	134	2,448	309	627,784	5,815,837	706,395
55-59	790	3,934	315	1,867,445	5,411,857	418,808
60-64	1,111	3,972	480	1,673,832	3,406,407	391,215
65-69	1,167	3,922	584	1,293,674	1,761,799	323,141
Female						
50-54	225	1,897	280	1,080,000	5,342,263	703,688
55-59	897	2,549	239	2,768,143	4,685,597	425,196
60-64	1,051	2,395	340	2,262,252	3,103,281	414,159
65-69	1,006	2,572	445	1,623,734	1,767,879	348,205
Total	6,381	23,689	2,992	13,196,864	31,294,920	3,730,807

Chiu H-M, *et al. Gut* 2021; 70:2321–2329. doi: 10.1136/gutjnl-2020-322545

Supplementary Table 3

Number of advanced CRC and the proportion of cancers with stage II or higher during the period of 1991 to 2000

Cohort / Dataset	N	Number of stage II+ colorectal cancer cases	%
NTUH	169	165	98
TVGH	3230	2746	85

NTUH: National Taiwan University Hospital

TVGH: Taipei Veteran General Hospital

The proportion of stage II and higher colorectal cancer was approximately 86% based on meta-analysis and Bayesian approach

Supplementary Table 4

Number of distal CRC and its proportion among all incident CRC in the period of 1991 to 2000

Cohort / Dataset	N	Number of distal colorectal cancers	%
NTUH	166	114	67
TVGH	3230	2552	79

NTUH: national Taiwan University Hospital

TVGH: Taipei Veteran General Hospital

The proportion of distal CRC was approximately 78% based on meta-analysis and Bayesian approach

Supplementary Table 5

Number of distal CRC death and its proportion among all CRC deaths in the period of 1981 to 2000

Cohort / Dataset	Colorectal cancer death, n	Distal colorectal cancer death, n	%
NTHU	105	70	67
TVGH	3143	2462	78

NTUH: national Taiwan University Hospital

TVGH: Taipei Veteran General Hospital

The proportion of death from distal site of CRC among all CRC death was approximately 77% based on meta-analysis and Bayesian approach

References

1. Wu CY, Anttila A, Yen AM, et al. Evaluation of breast cancer service screening programme with a Bayesian approach: mortality analysis in a Finnish region. Breast Cancer Res Treat. 2010;121:671-8.
2. Ju JH, Chang SC, Wang HS et al. Changes in disease pattern and treatment outcome of colorectal cancer: a review of 5,474 cases in 20 years. Int J Colorectal Dis. 2007;22:855-62.