# Statistical mediation analysis – overview, recent developments and an application in cancer epidemiology

Josef Fritz

Department of Medical Statistics and Informatics, Medical University of Innsbruck

Statistical mediation analyses provide tools to disentangle and assess the relative magnitude of different causal pathways and mechanisms by which a predictor (or exposure) may affect an outcome. The basic concepts of mediation analysis sound simple – but this is deceptive. Commonly occurring scenarios such as confounding, effect modification, categorical or time-to-event outcomes, and multiple mediators introduce the potential for biased results and/or pitfalls. Only over the last decade, new methods have been developed to resolve many of these problems satisfactorily. This was done by interpreting the problem of mediation in the framework of potential outcomes ('counterfactuals') and applying concepts of causal inference.

In my talk, I will give an overview of statistical methods for mediation analysis, from traditional regression-based approaches to state-of-the-art techniques. I will demonstrate concepts and discuss two concrete methods for effect estimation: (i) an extension of the traditional 'product method' for mediation (also allowing for exposure-mediator interaction) by VanderWeele, and (ii) natural effect models, a weighting and imputation-based approach for mediation, introduced by Lange et al. I will illustrate the application of these two methods using an example from my own research, where we investigated in how far the effect of body mass index on cancer incidence risk is mediated by insulin resistance.

References:

T. J. VanderWeele. Causal mediation analysis with survival data. Epidemiology. 2011

T. Lange et al. A Simple Unified Approach for Estimating Natural Direct and Indirect Effects. Am. J. Epidemiol. 2012

J. Fritz et al. The triglyceride-glucose index as a measure of insulin resistance and risk of obesity-related cancers. Int. J. Epidemiol. 2019

# Statistical mediation analysis – overview, recent developments and an application in cancer epidemiology

11/18/2020 – Virtual seminar at DCEG, Biostatistics Branch

**Josef Fritz**

josef.fritz@i-med.ac.at

Department of Medical Statistics, Informatics, and Health Economics

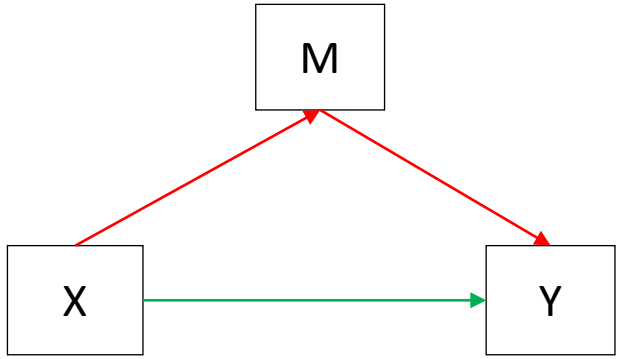
Medical University of Innsbruck

# Part 1: Mediation analysis from a causal inference perspective - Concepts, theory, and estimation methods

# Motivation

- Cause-and-effect relationship
- How does this effect come about?
- What are the underlying mechanisms?

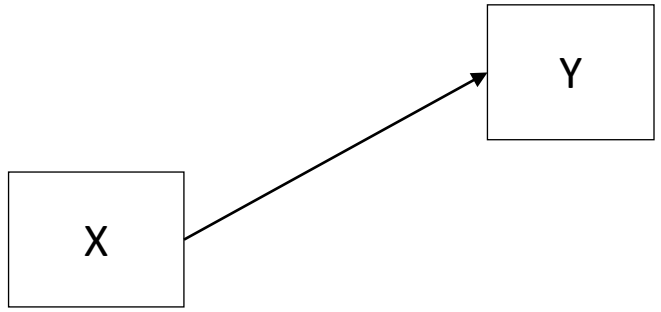
→Mediation analysis




- Statistical mediation analysis
  - Quantifying specific causal pathways measured via specific variables which are assumed to be affected by the exposure and themselves affect the outcome
  - Total effect
  - Indirect (mediated) effect
  - Direct effect

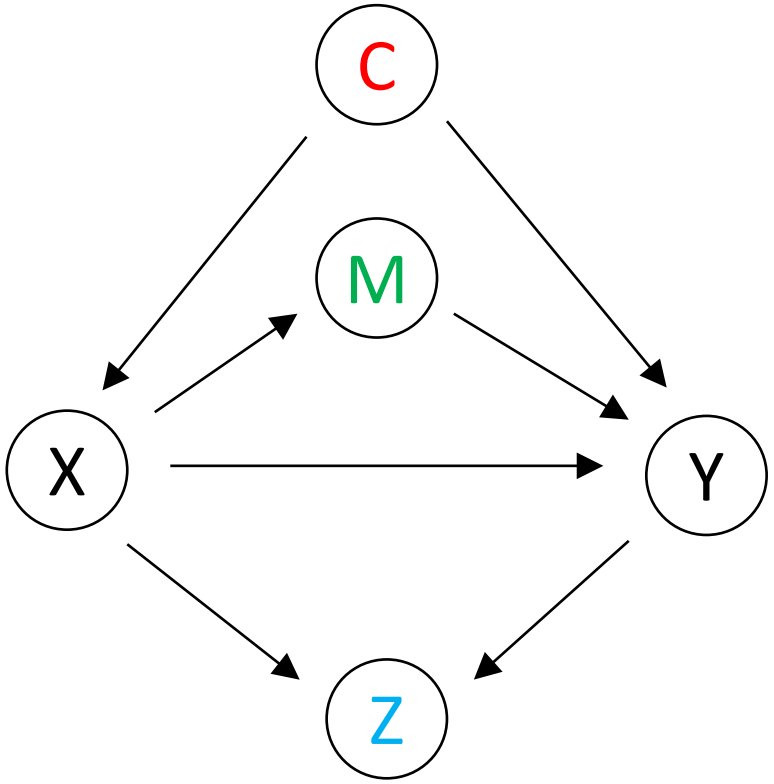
# Causal diagrams (directed acyclic graphs, DAGs)

- Illustrate the underlying causal assumptions
- Are based on knowledge in the respective research field
- An arrow indicates a causal effect of one variable on another
- A useful tool for epidemiologists
  - Selection of confounders
  - Graphical presentation of biases
  - Mediation analysis



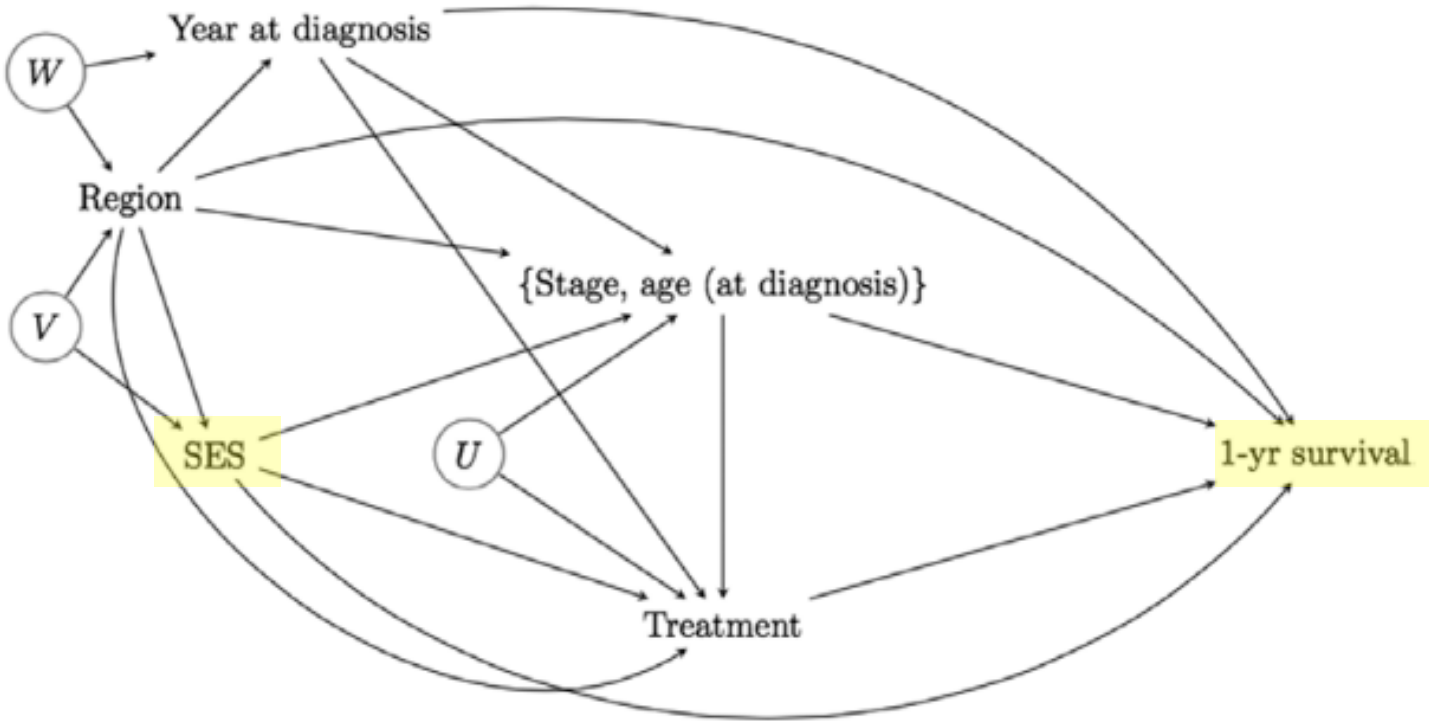
# Causal diagrams

- Path - a consecutive sequence of arrows
- Can be unblocked (open) or blocked
- An unblocked path can either represent a causal effect or transmit an association that is not causal (biasing path)




C – Confounder
M – Mediator
Z - Collider

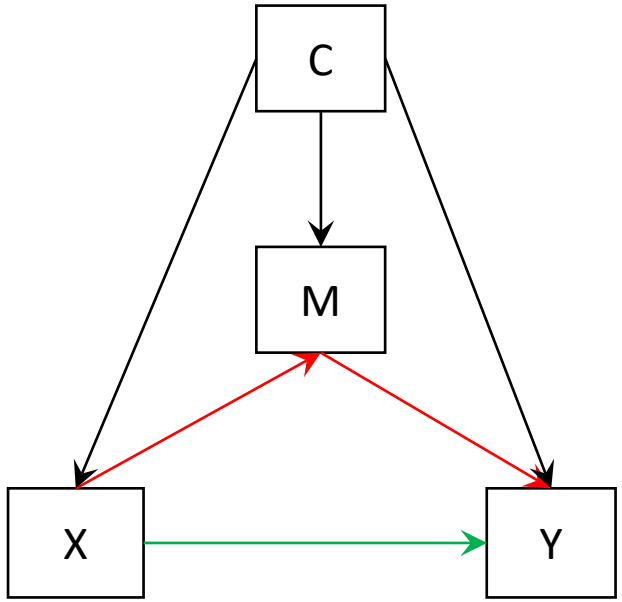
# Causal diagrams can be complex




Interventional Effects for Mediation Analysis with Multiple Mediators (*Epidemiology* 2017;28: 258–265)

*Stijn Vansteelandt*[a] *and Rhian M. Daniel*[b]

# Traditional difference method




1. Calculate a **model without the mediator**
   → **total effect**

$$E[Y|X,C] = \beta_0 + \beta_X X + \beta_C C$$

2. Calculate another **model conditioned on the mediator**
   → **direct effect**

$$E[Y|X,C,M] = \alpha_0 + \alpha_X X + \alpha_C C + \alpha_M M$$

3. If the coefficients $\beta_X$ and $\alpha_X$ differ, then some of the effect is thought to be mediated

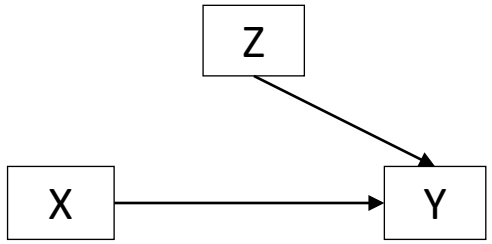
3'. The **indirect effect** is the **difference** between total and direct effect

T. VanderWeele. Mediation Analysis: A Practitioner's Guide. Annual review of public health, 2012.

# Challenges

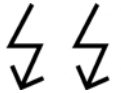
- Confounding
- Multiple mediators (parallel, sequential)
- Interactions between exposure and mediator(s)
- Non-linear dependencies

**→Need for generic, formal definitions of the desired effects (estimands).**

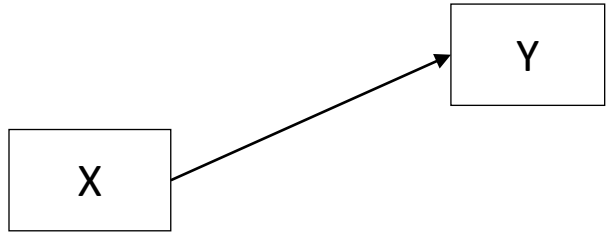



| | Z=1 (N=200) | | Z=0 (N=200) | | Marginal (N=400) | |
|---|---|---|---|---|---|---|
| | X=1 | X=0 | X=1 | X=0 | X=1 | X=0 |
| Y=1 | 90 | 70 | 30 | 10 | 120 | 80 |
| Y=0 | 10 | 30 | 70 | 90 | 80 | 120 |

- The conditional OR (“direct effect”) is 3.86 in both groups, but the marginal OR (“total effect”) is 2.25
- Difference method leads to wrong conclusion of an indirect effect being present!
- Non-collapsibility of the OR



# Counterfactual framework




- The potential outcome of a variable Y (outcome) is simply the value Y would have taken for individual u, had X (exposure) been assigned the value of x.
- In reality we are unable to see the outcome Y for an individual u given X or x, therefore we must rely on statistical modelling and assumptions for our estimates to be valid.
- J. Pearl, The Book of Why, 2018

# Potential outcomes and counterfactuals

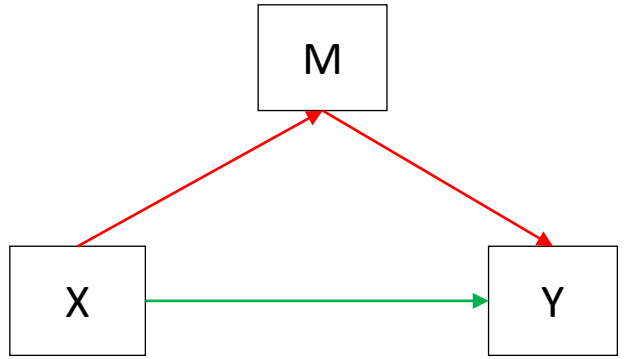
| | Flu shot (X) | Got the flu (Y) |
|---|---|---|
| Jonas | 0 | 1 |
| Josef | 1 | 0 |
| Mike | 1 | 0 |
| Sarah | 1 | 1 |
| Susan | 0 | 0 |
| Tom | 0 | 1 |

- Was the flu shot the cause why Josef did not get the flu?
- We do not know.
- To answer this we need to know what would have happened had he not gotten the flu shot.

- Y(X=0) and Y(X=1) represent the potential outcomes under both conditions.
- The outcome we do not observe is called the counterfactual.

# Total, natural direct and natural indirect effect

- **Total effect (TE)** of X on Y (level a vs. a*):
  - Y(X=a) - Y(X=a*) for each individual
  - E[Y(X=a)] – E[Y(X=a*)]
  - E[Y(X=a,M=M(a))] – E[Y(X=a*,M=M(a*))]



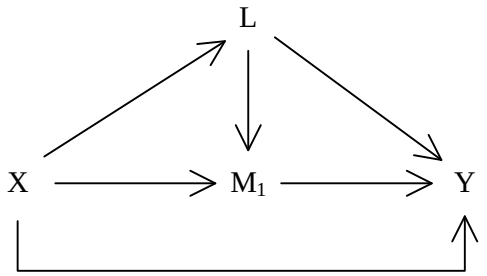

- **Natural indirect effect (NIE):**
  - E[Y(X=a,M=M(a))] – E[Y(X=a,M=M(a*))] (nested counterfactual!)
- **Natural direct effect (NDE):**
  - E[Y(X=a,M=M(a*))] – E[Y(X=a*,M=M(a*))]

→TE = NIE + NDE

- **Controlled direct effect (CDE):**
- E[Y(X=a,M=m)] – E[Y(X=a*,M=m)]

Pearl J. Direct and Indirect Effects. In: *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*. 2001.

# Mediation formula (Pearl, 2012)

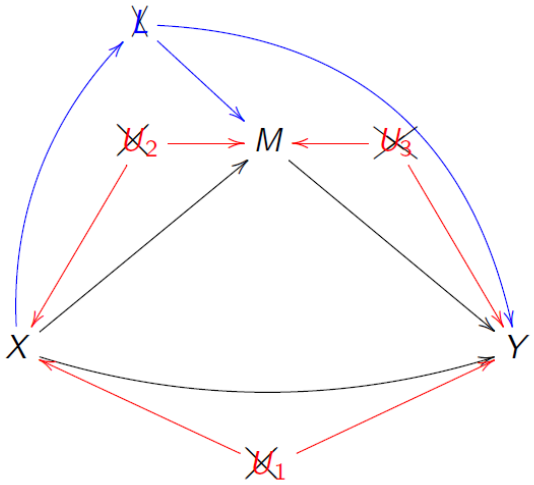



**IF** there is no uncontrolled confounding of the

- Exposure-outcome relationship
- The exposure-mediator relationship
- The mediator-outcome relationship

**AND IF** the 'cross-world' independence assumption holds, i.e. no exposure-induced mediator-outcome confounder

**AND IF** the positivity and consistency assumptions hold




**THEN** we have:

$$\text{E[Y(X=a,M=M}(a^*))] = \sum_{c \epsilon C, m \epsilon M} E[Y|X=a, M=m, C=c] * P(M=m|X=a^*, C=c) * P(C=c)$$

**non-parametric, curse of dimensionality → parametric models needed**

# Derivation of the mediation formula

Applied mediation analyses: a review and tutorial Epidemiology and Health, 2017

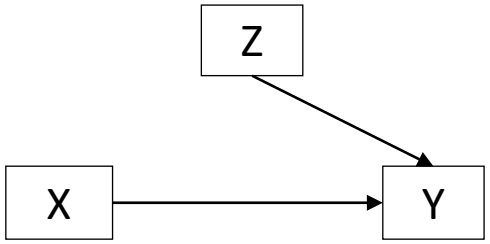
$$
\begin{aligned}
& E[g(Y(a, M(a^*)))] \\
= & \sum_{c \in \mathcal{C}, m \in \mathcal{M}} E[g(Y(a,m)) \mid M(a^*) = m, C = c] P(M(a^*) = m \mid C = c) P(C = c) \\
\overset{i}{=} & \sum_{c \in \mathcal{C}, m \in \mathcal{M}} E[g(Y(a,m)) \mid C = c] P(M(a^*) = m \mid C = c) P(C = c) \\
\overset{ii}{=} & \sum_{c \in \mathcal{C}, m \in \mathcal{M}} E[g(Y(a,m)) \mid A = a, M = m, C = c] P(M(a^*) = m \mid A = a^*, C = c) P(C = c) \\
\overset{iii}{=} & \sum_{c \in \mathcal{C}, m \in \mathcal{M}} E[g(Y) \mid A = a, M = m, C = c] P(M = m \mid A = a^*, C = c) P(C = c)
\end{aligned}
$$

i – cross-world independence assumption: $Y(a,m) \perp\!\!\!\perp M(a^*) \mid C$ for any $m$ and $a \neq a^*$

ii – no uncontrolled confounding : $Y(a, M(a)) \perp\!\!\!\perp A \mid C$ , $M(a) \perp\!\!\!\perp A \mid C$ , $Y(a,m) \perp\!\!\!\perp M \mid (A, C)$

iii – consistency: $P(Y(A,M) = Y) = 1$ and $P(M(A) = M) = 1$

# Mediation formula applied



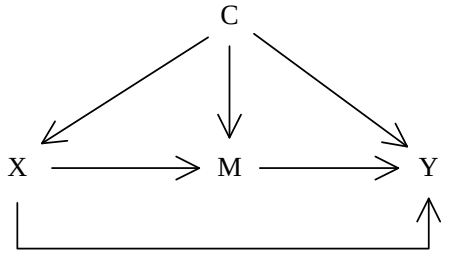

| | Z=1 (N=200) | | Z=0 (N=200) | | Marginal (N=400) | |
|---|---|---|---|---|---|---|
| | X=1 | X=0 | X=1 | X=0 | X=1 | X=0 |
| Y=1 | 90 | 70 | 30 | 10 | 120 | 80 |
| Y=0 | 10 | 30 | 70 | 90 | 80 | 120 |
| OR | 3.86 | | 3.86 | | 2.25 | |

E[Y(1,Z(1))] = 0.6 → Odds for Y(1,Z(1)) = 1.5
E[Y(1,Z(0))] = 0.6 → Odds for Y(1,Z(0)) = 1.5
E[Y(0,Z(0))] = 0.4 → Odds for Y(1,Z(0)) = 0.67

TE = Odds for Y(1,Z(1)) / Odds for Y(0,Z(0)) = 2.25
NIE = Odds for Y(1,Z(1)) / Odds for Y(1,Z(0)) = 1.0
NDE = Odds for Y(1,Z(0)) / Odds for Y(0,Z(0)) = 2.25
TE = NIE * NDE

# Regression-based approaches

Tyler Vanderweele




- Time-to-event outcome Y modeled via the following proportional hazards model (**"Outcome model"**):
  - ✓ $\lambda_Y(t|X=x,\mathrm{M}=\mathrm{m},C=c)=\lambda_Y(t|X=0,M=0,C=0)\times \exp(\alpha x+\beta m+\gamma xm+\delta c)$

- Continuous mediator via the following linear regression model (**"Mediator model"**)
  - ✓ $E(M|X=x,C=c)=\zeta+\eta x+\theta c$ (with error term $\delta^2$)

- Then, if the assumptions for the identification of NIEs and NDEs hold, and if the outcome is rare:
  - ✓ $NDE_{HR}(Y)_{x,x^*}=\exp(\alpha+\gamma\times(\zeta+\eta x^*+\theta \mathrm{c}+\beta\delta^2)\times(x-x^*)+0.5\gamma^2\delta^2\times\left(x^2-x^{*2}\right)$
  - ✓ $NIE_{HR}(Y)_{x,x^*}=\exp(\eta\beta+\eta\gamma x)\times(x-x^*)$

**R package regmedint (version 0.1.0)**

VanderWeele TJ. Causal mediation analysis with survival data. *Epidemiology,* 2012.
VanderWeele TJ. *Explanation in Causal Inference: Methods for Mediation and Interaction*. 1st ed. Oxford University Press, 2015.

# Natural effect models (NEMs)

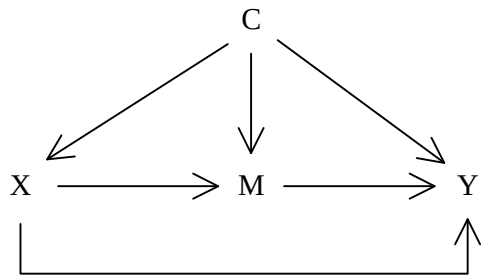
## Theis Lange (2012)

- NEM – a regression model for the nested counterfactual
  - $g(E[Y(a, M(a^*)]) = \alpha_0 + \alpha_1 * a + \alpha_2 * a^*$
  - g – link function
  - $\alpha_1$ – (marginal) natural direct effect
  - $\alpha_2$– (marginal) natural indirect effect

T. Lange et al. A simple unified approach for estimating natural direct and indirect effects. *Am J Epidemiol* 2012.
T. Lange et al. Assessing natural direct and indirect effects through multiple pathways. *Am J Epidemiol* 2014.

# Estimation algorithm for NEM (imputation-based approach)

1. Using the original data, fit a regression model to the outcome conditioned on confounders, exposure, and mediator (MODEL 1).




| Original dataset | | | | |
|---|---|---|---|---|
| ID | X | M | Y | C |
| 0001 | 0 | 1 | 0 | 0 |
| | | | | |
| 0002 | 1 | 1 | 1 | 0 |
| | | | | |
| 0003 | 0 | 0 | 1 | 1 |
| | | | | |
| Etc. | Etc. | Etc. | Etc. | Etc. |

X, M, Y, C all binary variables (0, 1)

# Estimation algorithm for NEM (imputation-based approach)

1. Using the original data, fit a regression model to the outcome conditioned on confounders, exposure, and mediator (MODEL 1).
2. Duplicate the original dataset, introducing a new auxiliary exposure variable (X*), which is equal to the original exposure for the first replication and equal to the opposite for the second replication.
3. Impute possible outcomes Y in the new rows.
   - ✓ Done using MODEL 1 from step 1.
4. Fit a regression model to the extended data set including X, X*, and C, but not the mediator.
   - ✓ Coefficient of X – NIE; coefficient of X* - NDE
5. Repeat (3) and (4) several times and take the average of the estimates.
6. CIs can be obtained using bootstrapping, repeating (1)-(5) 1000 times.

| Original dataset | | | | |
|---|---|---|---|---|
| ID | X | M | Y | C |
| 0001 | 0 | 1 | 0 | 0 |
| | | | | |
| 0002 | 1 | 1 | 1 | 0 |
| | | | | |
| 0003 | 0 | 0 | 1 | 1 |
| | | | | |
| Etc. | Etc. | Etc. | Etc. | Etc. |

X, M, Y, C all binary variables (0, 1)

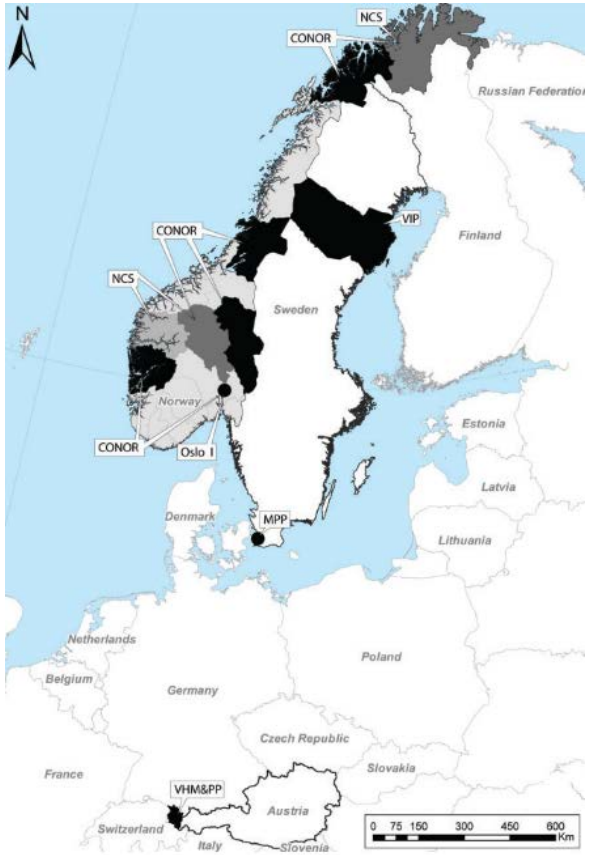
**R package medflex (version 0.6-7)**

# Part 2: Mediation analysis applied
# Does the TyG Index as a measure of insulin resistance mediate the effect of excess body weight on cancer risk?

# The Metabolic Syndrome and Cancer Project Me-Can project (Me-Can)

Me-Can

http://me-can.se/




- Pooling of six population-based European cohorts
- >800,000 individuals
- >80,000 incident cancers over a follow-up of ~20 years

- >30 original articles over the last 10 years

Cohort Profile: The Metabolic syndrome and Cancer project (Me-Can), Int J Epi, 2009

› Int J Epidemiol. 2020 Feb 1;49(1):193-204. doi: 10.1093/ije/dyz053.

# The triglyceride–glucose index as a measure of insulin resistance and risk of obesity–related cancers

Josef Fritz [1], Tone Bjørge [2 3], Gabriele Nagel [4 5], Jonas Manjer [6], Anders Engeland [2 7], Christel Häggström [8 9 10], Hans Concin [5], Stanley Teleka [11], Steinar Tretli [3], Björn Gylling [12], Alois Lang [5], Pär Stattin [9], Tanja Stocks [11], Hanno Ulmer [1]

Affiliations – collapse

## Affiliations

1 Department of Medical Statistics, Informatics and Health Economics, Medical University of Innsbruck, Innsbruck, Austria.
2 Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway.
3 Cancer Registry of Norway, Oslo, Norway.
4 Institute of Epidemiology and Medical Biometry, Ulm University, Ulm, Germany.
5 Agency for Preventive and Social Medicine, Bregenz (aks), Austria.
6 Department of Surgery, Skåne University Hospital, Lund University, Malmö, Sweden.
7 Division of Mental and Physical Health, Norwegian Institute of Public Health, Bergen, Norway.
8 Department of Biobank Research, Umeå University, Umeå, Sweden.
9 Department of Surgical Sciences, Uppsala University, Uppsala, Sweden.
10 Department of Public Health and Clinical Medicine, Nutritional Research, Umeå University, Umeå, Sweden.
11 Department of Clinical Sciences Lund, Lund University, Lund, Sweden.
12 Department of Medical Biosciences, Pathology, Umeå University, Umeå, Sweden.

PMID: 30945727 DOI: 10.1093/ije/dyz053

Hanno Ulmer

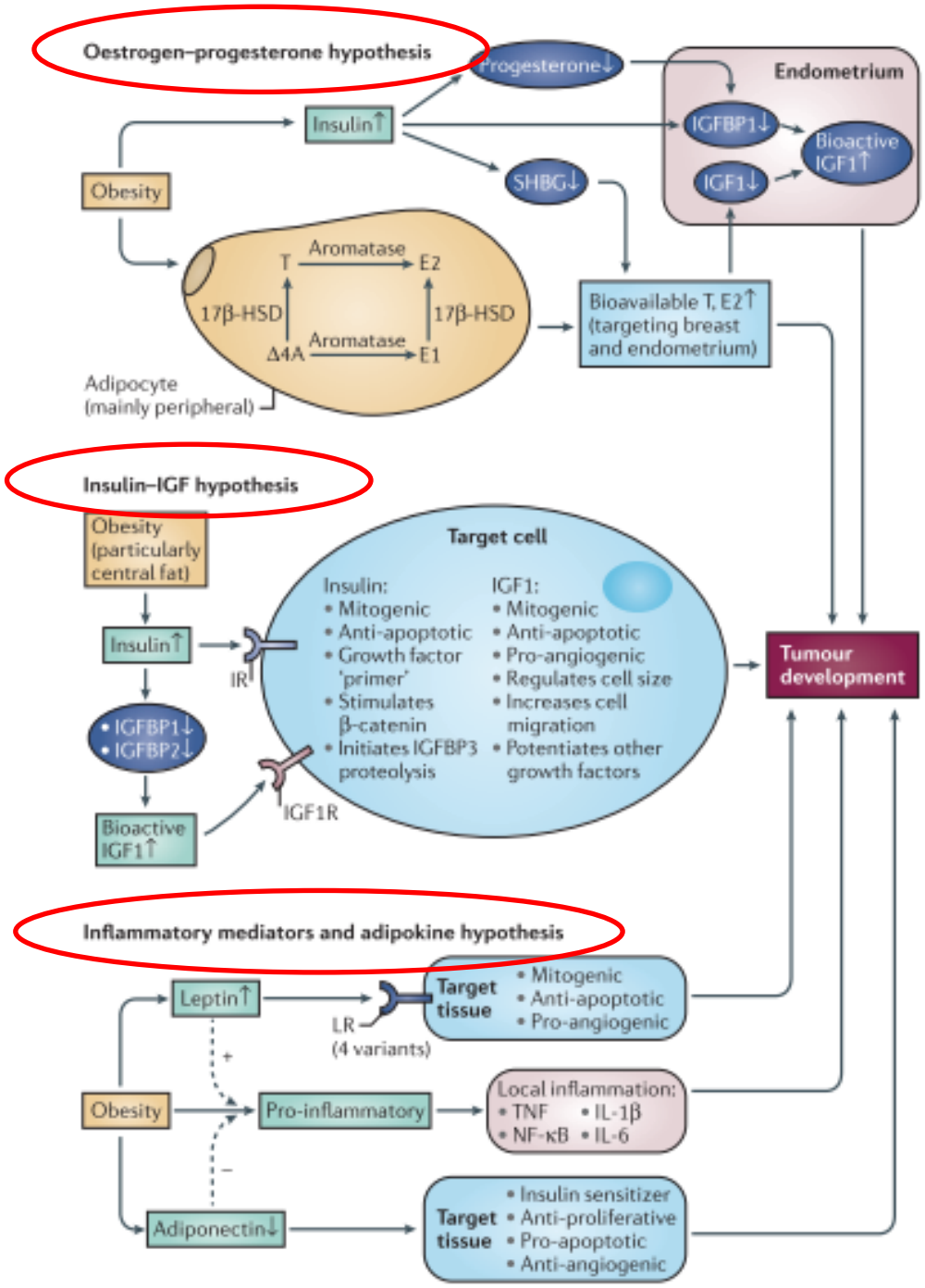
Tanja Stocks

# Excess body weight is a major risk factor for many cancer forms

**Table 2. Strength of the Evidence for a Cancer-Preventive Effect of the Absence of Excess Body Fatness, According to Cancer Site or Type.***

| Cancer Site or Type | Strength of the Evidence in Humans† | Relative Risk of the Highest BMI Category Evaluated versus Normal BMI (95% CI)‡ |
|---|---|---|
| Esophagus: adenocarcinoma | Sufficient | 4.8 (3.0–7.7) |
| Gastric cardia | Sufficient | 1.8 (1.3–2.5) |
| Colon and rectum | Sufficient | 1.3 (1.3–1.4) |
| Liver | Sufficient | 1.8 (1.6–2.1) |
| Gallbladder | Sufficient | 1.3 (1.2–1.4) |
| Pancreas | Sufficient | 1.5 (1.2–1.8) |
| Breast: postmenopausal | Sufficient | 1.1 (1.1–1.2)§ |
| Corpus uteri | Sufficient | 7.1 (6.3–8.1) |
| Ovary | Sufficient | 1.1 (1.1–1.2) |
| Kidney: renal-cell | Sufficient | 1.8 (1.7–1.9) |
| Meningioma | Sufficient | 1.5 (1.3–1.8) |
| Thyroid | Sufficient | 1.1 (1.0–1.1)§ |
| Multiple myeloma | Sufficient | 1.5 (1.2–2.0) |

Body Fatness and Cancer — Viewpoint of the IARC Working Group : New Engl J Med, 2016




# Excess body weight and cancer risk: proposed biological mechanisms

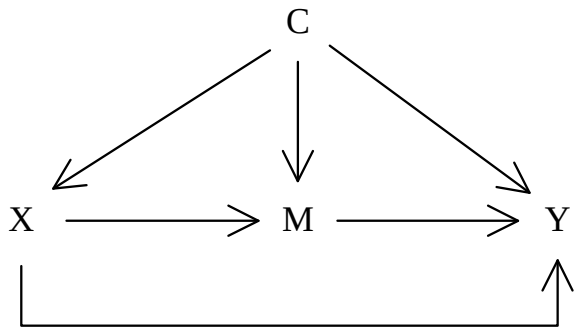
Renehan AG, Zwahlen M, Egger M. Adiposity and cancer risk: new mechanistic insights from epidemiology. *Nat Rev Cancer* 2015.

# TyG Index: A Novel Measure for Insulin Resistance

- TyG index = the logarithmized product of fasting levels of triglycerides and glucose (“TyG index”).
- Simple measure of insulin resistance.
- Highly correlated with the euglycemic-hyperinsulinemic clamp test (Pearson correlation of -0.68), similar magnitude as the frequently used HOMA-IR.
- A promising surrogate measure for insulin resistance in large-scale epidemiological studies.

- First proposed by: Simental-Mendía LE et al. The Product of Fasting Glucose and Triglycerides As Surrogate for Identifying Insulin Resistance in Apparently Healthy Subjects. *Metab Syndr Relat Disord* 2008.

# Study design & Methods




- Prospective cohort study (cohorts from Austria, Norway, and Sweden)
- 510,471 participants free of cancer at baseline
- Mean age at baseline: 43.1 years
- Median follow-up time: 17.2 years
- X: BMI as a measure of excess body weight, at baseline
- M: TyG index as a measure of insulin resistance, at baseline
- Y: incident cancer, various sites
- C: baseline age, sex, smoking status, fasting status, cohort, decade of birth
- First 12 month of follow-up excluded
- Regression-based approach from VanderWeele with exposure-mediator interaction (see slide no. 15)

# Baseline characteristics by quintiles of TyG index, Me-Can population

| | TyG index | | | | |
|---|---|---|---|---|---|
| | **Quintile 1 (N=102,521)** | **Quintile 2 (N=102,020)** | **Quintile 3 (N=101,851)** | **Quintile 4 (N=101,954)** | **Quintile 5 (N=102,125)** |
| **TyG index, mean (SD)** | 7.8 (0.2) | 8.3 (0.1) | 8.5 (0.1) | 8.9 (0.1) | 9.5 (0.4) |
| **TyG index, range** | <8.1 | 8.1 to 8.4 | 8.4 to 8.7 | 8.7 to 9.1 | >9.1 |
| **BMI categories** | | | | | |
| **<18.5 kg/m2** | 3.8% | 2.1% | 1.3% | 0.7% | 0.3% |
| **18.5 to 24.9 kg/m2** | 72.1% | 61.7% | 53.3% | 42.6% | 28.7% |
| **25 to 29.9 kg/m2** | 20.7% | 29.7% | 35.6% | 42.4% | 48.7% |
| **≥30.0 kg/m2** | 3.4% | 6.5% | 9.7% | 14.2% | 22.3% |
| **Sex, male** | 32.3% | 40.5% | 48.7% | 58.7% | 72.5% |
| **Age, yrs, mean (SD)** | 39.6 (11.0) | 42.8 (10.7) | 43.6 (10.7) | 44.4 (10.2) | 44.9 (9.4) |

HRs of BMI and TyG index on cancer risk from a standard multivariable Cox model (conditioned on BMI, TyG index, baseline age, sex, smoking status, fasting status, cohort, and decade of birth)

Total (in green) and indirect (through TyG index; in red) effects of BMI on cancer risk from the causal model with confounders baseline age, sex, smoking status, fasting status, cohort, and decade of birth. Bootstrap 95% CIs.

| Cancer site (ICD-7; ICD-10) | N of cases |
|---|---|
| **Endometrium (172; C54)** | 1,417 |
| **Liver (155.0; C22)** | 561 |
| **Oesophagus (adenocarcinoma) (150; C15)** | 185 |
| **Kidney (renal cell) (180.0, 180.9; C64)** | 1,347 |
| **Gallbladder (155.1-155.3; C23-24)** | 364 |
| **Colon (153; C18)** | 4,032 |
| **Pancreas (157; C25)** | 1,368 |
| **Rectum (154; C19-21)** | 2,430 |
| **Breast (postmenopausal) (170; C50)** | 3,427 |
| **Ovary (175.0; C56)** | 921 |
| **Gastrointestinal cancers combined** | 8,940 |
| **Gynecological cancers combined** | 5,765 |

# The TyG Index substantially mediated the effect of BMI for gastrointestinal cancers, but not for gynecological cancers.

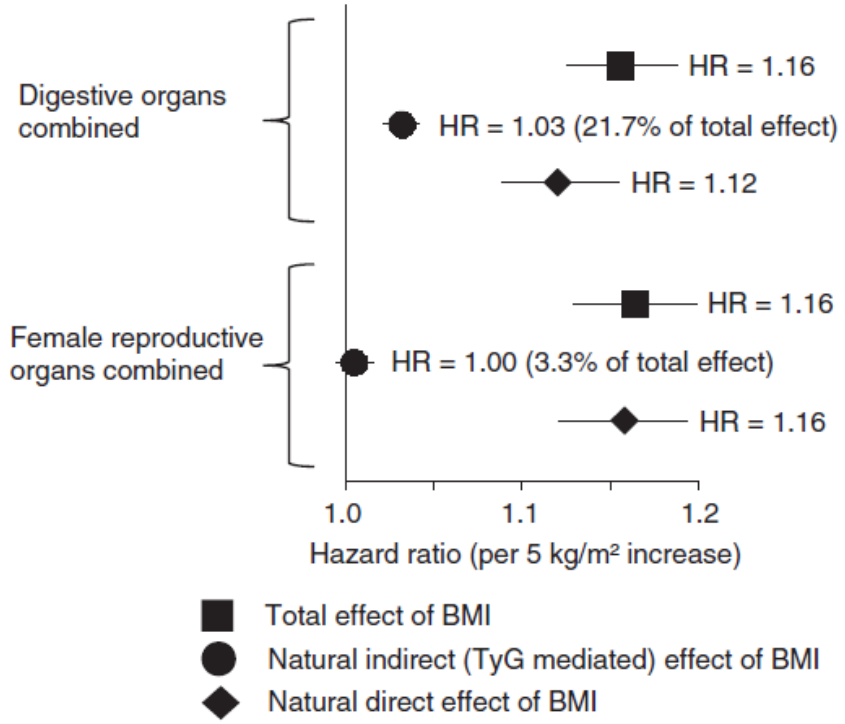



**Figure 1**. Total, natural indirect (TyG mediated) and direct effects of continuous BMI on cancer risk of digestive organs (oesophagus, colon, rectum, liver, gallbladder and pancreas) vs female reproductive organs [endometrium, ovary and breast (postmenopausal)].

- ✓ Our results confirm a promoting role of insulin resistance as measured through the TyG index in the pathogenesis of gastrointestinal cancers.
- ✓ Our results provide limited evidence that insulin resistance as measured through the TyG index connects excess body weight with risk of cancers of the female reproductive organs.

# Summary

- Causal statistical mediation analysis in the counterfactual framework
- Methods of estimates
- Application on an oncological research question
  - ✓ Role of insulin resistance in the BMI – cancer risk relationship

**<u>Current/Ongoing work:</u>**

- Effect decomposition in the case of multiple mediators potentially causally affecting each other
- Single and joint contributions of the pathways through

  (i) insulin resistance, (ii) hypertension, (iii) hyperuricemia and (iv) hypercholesterolemia

  to the total effect of obesity on end-stage kidney disease

# Any questions?

# Back-up slides

# Remarks - regression-based approaches

Tyler Vanderweele

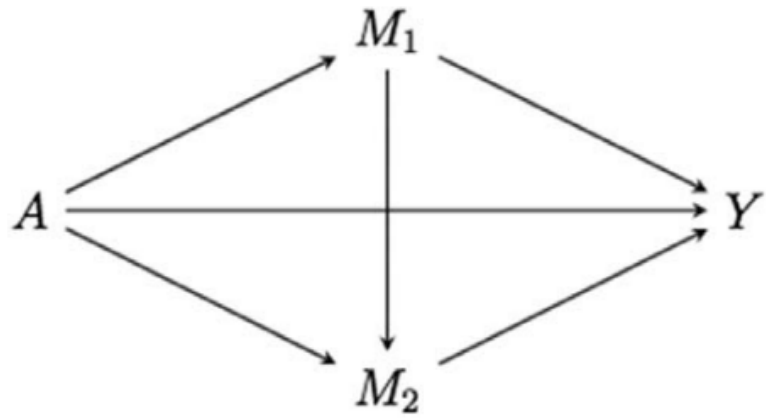
- Derivation of NIE and NDE for various outcome and mediator models
  - ✓ Outcome: additive hazards model, accelerated failure time model
  - ✓ Mediator: continuous, binary
- New formulas each time
- In case of linear models and no exposure-mediator interaction, NIE und NDE simplify to the path-specific effects from structural equation models (SEMs)
- Extensions for several mediators ‘en bloc’
- R package regmedint (version 0.1.0)

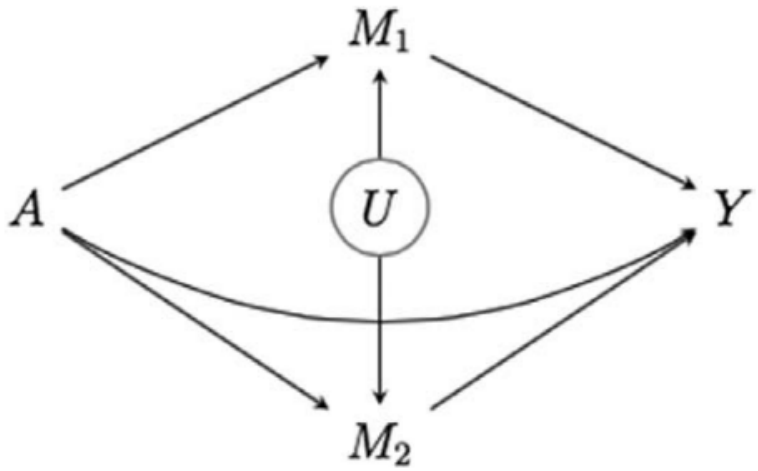
# Remarks - Natural effect models (NEMs)

## Theis Lange (2012)

- Flexible, since the two regression models involved (step 1 and 4) can be of any type
- Mediator can also be a vector of variables
- Extensions for continuous exposures
- Weighting-based alternative
- Generalization to multiple mediators (not causally affecting each other)
- R package medflex (version 0.6-7)

# Extensions: Multiple mediators




- Beyond natural indirect effects – decomposition of total effect into direct effect, indirect effect through $M_1$, and indirect effect through $M_2$

→ Sequential mediation analysis

- ✓ VanderWeele TJ, Vansteelandt S. Mediation Analysis with Multiple Mediators. Epidemiol Method, 2014.

→ **Interventional direct and indirect effects**

- ✓ Vansteelandt S, Daniel RM. Interventional Effects for Mediation Analysis with Multiple Mediators. Epidemiology, 2017.

# Extensions: Time-varying exposures and mediators

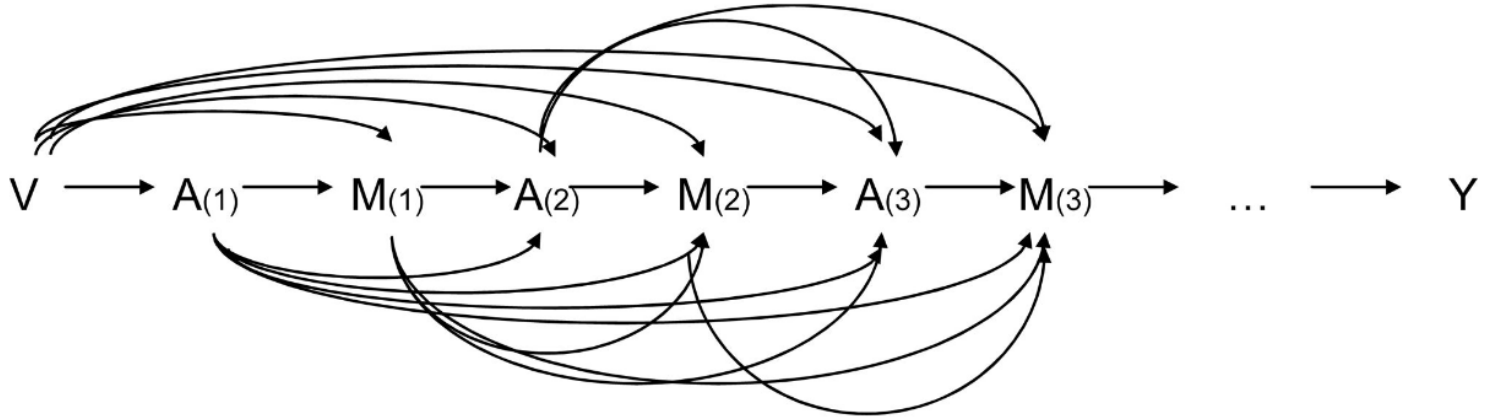



**Figure 4.**
Time-varying exposures and mediators, with no time-varying confounders.

*J R Stat Soc Series B Stat Methodol.* 2017 June ; 79(3): 917–938. doi:10.1111/rssb.12194.

**Mediation analysis with time varying exposures and mediators**

**Tyler J. VanderWeele** and **Eric J. Tchetgen Tchetgen**

**Table 1.** Baseline characteristics, Me-Can 2.0 study population

| Characteristic | *n* (%) resp. mean (SD), median |
|---|---|
| Cohort (year of baseline measurement) | |
| Oslo study I (1972-73) | 17 644 (3.5%) |
| NCS (1974-88) | 61 209 (12.0%) |
| 40y (1985-99) | 134 519 (26.4%) |
| VHM&PP (1985-2005) | 173 538 (34.0%) |
| VIP (1985-2014) | 92 995 (18.2%) |
| MPP (1974-2006) | 30 566 (6.0%) |
| Total | 510 471 (100%) |
| Sex | |
| Male | 257 968 (50.5%) |
| Female | 252 503 (49.5%) |
| Baseline age, years | 43.1 (10.6), 41.5 |
| Smoking status | |
| Never smoker | 241 940 (47.4%) |
| Ex-smoker | 136 417 (26.7%) |
| Current smoker | 132 114 (25.9%) |
| Decade of birth | |
| ≤1929 | 51 894 (10.2%) |
| 1930-39 | 85 098 (16.7%) |
| 1940-49 | 77 228 (15.1%) |
| 1950-59 | 197 575 (38.7%) |
| 1960-69 | 66 238 (13.0%) |
| ≥1970 | 32 438 (6.4%) |
| BMI, $kg/m^2$ | 25.2 (4.0), 24.7 |
| BMI categories | |
| <18.5 $kg/m^2$ | 8355 (1.6%) |
| 18.5 to 24.9 $kg/m^2$ | 264 012 (51.7%) |
| 25 to 29.9 $kg/m^2$ | 180 896 (35.4%) |
| ≥30.0 $kg/m^2$ | 57 208 (11.2%) |
| Fasting status | |
| Less than 8 h | 210 350 (41.2%) |
| 8 h or more | 300 121 (58.8%) |
| Glucose, mmol/l | 5.26 (1.22), 5.14 |
| Fasting (8 h or more) individuals | 5.14 (1.21), 5.05 |
| Triglycerides, mmol/l | 1.56 (1.10), 1.26 |
| Fasting (8 h or more) individuals | 1.43 (1.03), 1.16 |
| TyG index[a] | 8.60 (0.60), 8.55 |
| Fasting (8 h or more) individuals | 8.50 (0.58), 8.44 |

Oslo, Oslo study I; NCS, Norwegian Counties Study; 40-y, 40-year programme; VHM&PP, Vorarlberg Health Monitoring and Prevention Programme; VIP, Västerbotten Intervention Programme; MPP, Malmö Preventive Project.

[a]TyG index calculated as ln[triglycerides (mg/dl) x blood glucose (mg/dl)/2].

**Table 2.** Baseline characteristics by quintiles of TyG index

| Characteristic[a] | TyG index | | | | |
|---|---|---|---|---|---|
| | Quintile 1 ($N$ = 102 521) | Quintile 2 ($N$ = 102 020) | Quintile 3 ($N$ = 101 851) | Quintile 4 ($N$ = 101 954) | Quintile 5 ($N$ = 102 125) |
| TyG index[b] | 7.8 (0.2), 7.9 | 8.3 (0.1), 8.2 | 8.5 (0.1), 8.5 | 8.9 (0.1), 8.9 | 9.5 (0.4), 9.4 |
| TyG index, range[b] | <8.1 | 8.1 to 8.4 | 8.4 to 8.7 | 8.7 to 9.1 | >9.1 |
| BMI, kg/m$^2$ | 23.2 (3.2), 22.8 | 24.3 (3.6), 23.8 | 25.1 (3.8), 24.6 | 26.1 (4.0), 25.6 | 27.4 (4.2), 26.9 |
| BMI categories | | | | | |
| <18.5 kg/m$^2$ | 3859 (3.8%) | 2117 (2.1%) | 1359 (1.3%) | 731 (0.7%) | 289 (0.3%) |
| 18.5 to 24.9 kg/m$^2$ | 73 921 (72.1%) | 62 974 (61.7%) | 54 334 (53.3%) | 43 441 (42.6%) | 29 342 (28.7%) |
| 25 to 29.9 kg/m$^2$ | 21 268 (20.7%) | 30 299 (29.7%) | 36 304 (35.6%) | 43 256 (42.4%) | 49 769 (48.7%) |
| ≥30.0 kg/m$^2$ | 3473 (3.4%) | 6630 (6.5%) | 9854 (9.7%) | 14 526 (14.2%) | 22 725 (22.3%) |
| Sex | | | | | |
| Male | 33 153 (32.3%) | 41 317 (40.5%) | 49 640 (48.7%) | 59 811 (58.7%) | 74 047 (72.5%) |
| Female | 69 368 (67.7%) | 60 703 (59.5%) | 52 211 (51.3%) | 42 143 (41.3%) | 28 078 (27.5%) |
| Baseline age, years | 39.6 (11.0), 40.2 | 42.8 (10.7), 41.2 | 43.6 (10.7), 41.7 | 44.4 (10.2), 42.1 | 44.9 (9.4), 42.2 |
| Smoking status | | | | | |
| Never smoker | 60 550 (59.1%) | 54 606 (53.5%) | 48 131 (47.3%) | 42 555 (41.7%) | 36 098 (35.3%) |
| Ex-smoker | 22 541 (22.0%) | 24 681 (24.2%) | 26 723 (26.2%) | 29 507 (28.9%) | 32 965 (32.3%) |
| Current smoker | 19 430 (19.0%) | 22 733 (22.3%) | 26 997 (26.5%) | 29 892 (29.3%) | 33 062 (32.4%) |
| Fasting status | | | | | |
| Less than 8 h | 29 117 (28.4%) | 33 518 (32.9%) | 41 691 (40.9%) | 48 353 (47.4%) | 57 671 (56.5%) |
| 8 h or more | 73 404 (71.6%) | 68 502 (67.1%) | 60 160 (59.1%) | 53 601 (52.6%) | 44 454 (43.5%) |

[a]Given as $n$ (%) resp. mean (SD), median (except TyG index, range).

[b]TyG index calculated as ln[triglycerides (mg/dl) x blood glucose (mg/dl)/2].

**Table 4.** Decomposition of the total effect of continuous BMI on cancer risk into natural direct and indirect effect mediated by the TyG index, stratified by cancer site

| Site (ICD-7; ICD-10) | Total effect[a] HR (95% CI) | Natural indirect effect[a] HR (95% CI) | Natural direct effect[a] HR (95% CI) | Proportion mediated (95% CI) |
|---|---|---|---|---|
| Oesophagus (adenocarcinoma[b]) (150; C15) | 1.48 (1.23 to 1.73) | 1.03 (0.96 to 1.08) | 1.44 (1.20 to 1.70) | 6.5% (−10.8% to 24.2%) |
| Colon (153; C18) | 1.14 (1.10 to 1.19) | 1.03 (1.01 to 1.04) | 1.11 (1.07 to 1.16) | 19.9% (9.4% to 35.1%) |
| Rectum (154; C19-21) | 1.09 (1.03 to 1.15) | 1.03 (1.01 to 1.05) | 1.06 (1.00 to 1.12) | 33.9% (11.8% to 100%) |
| Liver (155.0; C22) | 1.48 (1.33 to 1.63) | 1.04 (1.01 to 1.08) | 1.42 (1.27 to 1.57) | 11.1% (1.7% to 21.7%) |
| Gallbladder (155.1-155.3; C23-24) | 1.30 (1.14 to 1.46) | 1.04 (0.99 to 1.09) | 1.25 (1.09 to 1.42) | 15.9% (−2.6% to 44.0%) |
| Pancreas (157; C25) | 1.11 (1.03 to 1.20) | 1.05 (1.02 to 1.07) | 1.06 (0.99 to 1.15) | 41.7% (16.0% to 100%) |
| Pancreas, males | 1.17 (1.05 to 1.31) | 1.03 (0.99 to 1.07) | 1.14 (1.01 to 1.29) | 19.9% (−4.9% to 81.4%) |
| Pancreas, females | 1.07 (0.97 to 1.17) | 1.06 (1.02 to 1.09) | 1.01 (0.90 to 1.11) | 90.8% (−100% to 100%) |
| Breast (postmenopausal) (170; C50) | 1.05 (1.01 to 1.09) | 1.01 (0.99 to 1.02) | 1.04 (1.00 to 1.09) | 15.9% (−18.0% to 79.5%) |
| Endometrium (172; C54) | 1.50 (1.43 to 1.57) | 1.01 (0.99 to 1.03) | 1.49 (1.41 to 1.57) | 1.6% (−3.5% to 6.5%) |
| Ovary (175.0; C56) | 1.05 (0.96 to 1.13) | 0.99 (0.97 to 1.02) | 1.06 (0.97 to 1.15) | −14.4% (−100% to 100%) |
| Kidney (renal cell) (180.0, 180.9; C64) | 1.36 (1.27 to 1.44) | 1.05 (1.02 to 1.07) | 1.30 (1.21 to 1.38) | 14.7% (6.6% to 23.6%) |
| Digestive organs combined[c] | 1.16 (1.13 to 1.19) | 1.03 (1.02 to 1.04) | 1.12 (1.09 to 1.16) | 21.7% (14.6% to 30.6%) |
| Endometrium, ovary and breast (postmenopausal) combined | 1.16 (1.13 to 1.20) | 1.00 (1.00 to 1.02) | 1.16 (1.12 to 1.19) | 3.3% (-3.3% to 10.8%) |

[a]HRs (per 5-kg/m$^2$ increase) were estimated according to the two-stage regression method proposed by VanderWeele[24], adjusted for baseline age, sex, smoking status, fasting status, cohort and decade of birth, with attained age as the underlying time scale. Analyses were restricted to participants with a BMI $\geq$ 18.5 (i.e. no underweight).

[b]Adenocarcinomas were identified via information on morphology (ICD-O-3 morphological key).

[c]Digestive organs combined include the following sites: oesophagus (adenocarcinoma), colon, rectum, liver, gallbladder and pancreas.

## Table S5: Associations of BMI with cancer risk, with and without adjusting for TyG index, stratified by cancer site

| Site (ICD-7; ICD-10) | Adjusted HR (95% CI)† Model 1 | Adjusted HR (95% CI)† Model 2 (with TyG index) |
|---|---|---|
| **Oesophagus (adenocarcinoma‡) (150; C15)** | 1.48 (1.22 to 1.78) | 1.44 (1.18 to 1.76) |
| **Colon (153; C18)** | 1.14 (1.09 to 1.19) | 1.11 (1.06 to 1.16) |
| **Rectum (154; C19-21)** | 1.09 (1.04 to 1.15) | 1.06 (1.00 to 1.12) |
| **Liver (155.0; C22)** | 1.47 (1.34 to 1.62) | 1.42 (1.28 to 1.57) |
| **Gallbladder (155.1-155.3; C23-24)** | 1.29 (1.15 to 1.46) | 1.25 (1.09 to 1.42) |
| **Pancreas (157; C25)** | 1.11(1.04 to 1.19) | 1.06 (0.99 to 1.15) |
| **Pancreas - males** | 1.17 (1.05 to 1.31) | 1.14 (1.01 to 1.28) |
| **Pancreas - females** | 1.07 (0.97 to 1.17) | 1.01(0.91 to 1.11) |
| **Breast (postmenopausal) (170; C50)** | 1.05 (1.01 to 1.09) | 1.04 (1.00 to 1.09) |
| **Endometrium (172; C54)** | 1.50 (1.43 to 1.57) | 1.49 (1.41 to 1.57) |
| **Ovary (175.0; C56)** | 1.05 (0.97 to 1.13) | 1.06 (0.97 to 1.15) |
| **Kidney (renal cell) (180.0, 180.9; C64)** | 1.35 (1.27 to 1.44) | 1.30 (1.21 to 1.39) |
| **Digestive organs combined§** | 1.16 (1.12 to 1.19) | 1.12 (1.09 to 1.15) |
| **Endometrium, ovary and breast (postmenopausal) combined** | 1.16 (1.13 to 1.20) | 1.16 (1.12 to 1.19) |

Model 1: adjusted for baseline age, sex, smoking status, fasting status, cohort, and decade of birth.

Model 2: adjusted for the same variables as in Model 1, plus additionally for TyG index.