Estimating breast cancer incidence using multiple imputation with chained equations (MICE)

Anna Johansson

anna.johansson@ki.se

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Outline

- Brief background: Breast Cancer and its subtypes
- Motivation and aim: Estimating subtype-specific BC incidence using cancer registry data
- Available data – partly missing information on subtype
- Incidence estimation using multiple imputation with chained equations (MICE)
  - → Why this is a non-standard multiple imputation
- Conclusions

This work was recently published

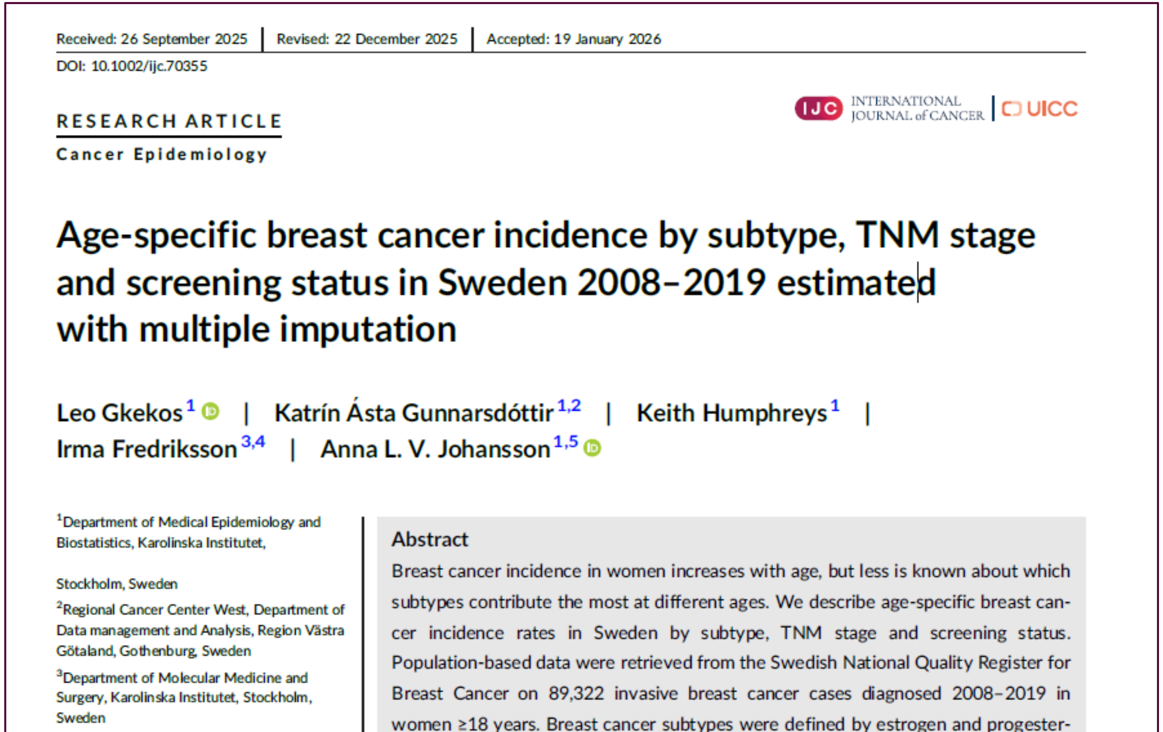
Int J Cancer. 2026 Feb 12.

PMID: 41676860

DOI: 10.1002/ijc.70355

Methods described in Supplemental material.

https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.70355



Received: 26 September 2025 | Revised: 22 December 2025 | Accepted: 19 January 2026

DOI: 10.1002/ijc.70355

IJC INTERNATIONAL JOURNAL of CANCER | UICC

RESEARCH ARTICLE

Cancer Epidemiology

Age-specific breast cancer incidence by subtype, TNM stage and screening status in Sweden 2008–2019 estimated with multiple imputation

Leo Gkekos[1] | Katrín Ásta Gunnarsdóttir[1,2] | Keith Humphreys[1] | Irma Fredriksson[3,4] | Anna L. V. Johansson[1,5]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet,

Stockholm, Sweden

[2]Regional Cancer Center West, Department of Data management and Analysis, Region Västra Götaland, Gothenburg, Sweden

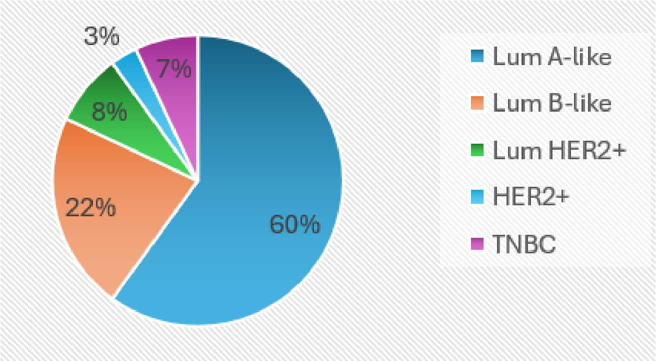
[3]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden

Abstract

Breast cancer incidence in women increases with age, but less is known about which subtypes contribute the most at different ages. We describe age-specific breast cancer incidence rates in Sweden by subtype, TNM stage and screening status. Population-based data were retrieved from the Swedish National Quality Register for Breast Cancer on 89,322 invasive breast cancer cases diagnosed 2008–2019 in women ≥18 years. Breast cancer subtypes were defined by estrogen and progester-

Background

- Breast cancer is not one disease but many **subtypes** with different treatment and survival.
  - Underlying **molecular subtypes** are approximated by clinical subtype ("**surrogate subtype**").
  - **5 "surrogate subtypes**": ER, PR, HER2 receptors, grade and KI67.



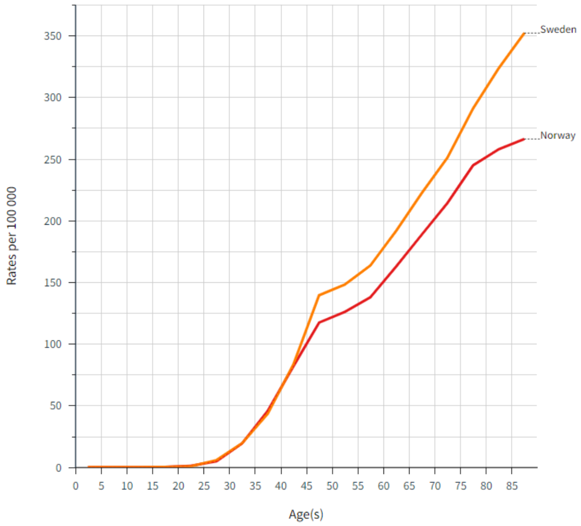
- Breast cancer subtypes have (partly) different underlying **risk factors** and **age-specific risk**.
- Also, **screening** is more likely to detect luminal tumours, while interval cancers are more often TNBC.

Screening has changed the age-specific incidence pattern of breast cancer

- Data from NORDCAN.

Rates per 100 000, Incidence, Females, [1975-1984]
Breast
Norway - Sweden

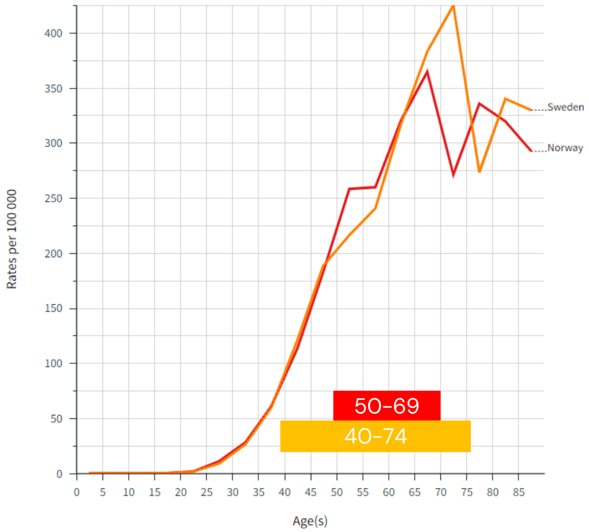
Before screening



NORDCAN | **IARC** - https://gco.iarc.who.int
Data version: 9.4 - 07.2024
© All Rights Reserved 2024

Rates per 100 000, Incidence, Females, [2013-2022]
Breast
Norway - Sweden

After screening



NORDCAN | **IARC** - https://gco.iarc.who.int
Data version: 9.4 - 07.2024
© All Rights Reserved 2024

Our aim: Incidence rates of BC subtypes across age

- **Our aim** was to estimate and describe **subtype-specific BC incidence across age**.

- This will be correlated with screening, as well as shifts in underlying risk factors.
  - E.g. reproductive risk factors, obesity, alcohol, MHT (menopausal hormone therapy), genetics, etc.

- These results are important as basis for studies on clinical outcomes, e.g. understanding case-mix over time and ages, and studies on screening, e.g. extended screening ages.

What data did we have

- Swedish National Quality Register of Breast Cancer (NKBC) was established in 2008.
  - Nationwide (population-based) cancer register.
  - Collects detailed **routine clinical** information on **all new** breast cancer cases diagnosed in women and men.
    - Patient and **tumour characteristics**, diagnostic work-up, planned/given treatment, follow-up.
  - One recorded tumour per side (max 2 per person) for Swedish residents only.

- We had data for women diagnosed with **invasive** BC 2008-2019, aged ≥20 at diagnosis. N=89,322

- Information on
  - **Date and age of diagnosis**
  - **Subtype: ER, PR, HER2, grade**
  - Stage: tumour size (T), lymph node metastasis (N), distant metastasis (M)
  - Screen-detected or symptomatically detected

What data did we have

- 5 surrogate subtypes were based on: ER, PR, HER2, grade.

Surrogate subtype	ER	PR	HER2	Grade
Luminal A-like	+ + –	+ – +	– – –	Grade 1,2 Grade 1 Grade 1,2
Luminal B-like	+ + –	+ – +	– – –	Grade 3 Grade 2,3 Grade 3
Luminal HER2	+	+/-	+	Grade 1-3
HER2 positive	–	–	+	Grade 1-3
Triple-negative (TNBC)	–	–	–	Grade 1-3
% missing	4.1%	4.2%	7.9%	16.4%

15.5% missing subtype

Incidence rate estimation

- Incidence rate: $\lambda(\text{age, year}) = N_{\text{cases}} / N_{\text{poprisk}}$
- N_{cases} : From NKBC : **case population (individual level)**
- N_{poprisk} : From Statistics Sweden: **population at risk (counts** by 1-year age, 1-year calendar time)
- Estimated rates separately for each of the 5 subtypes.
  - Lum A: $\lambda_1(\text{age, year}) = N_{\text{cases, subtype 1}} / N_{\text{poprisk}}$
  - Lum B: $\lambda_2(\text{age, year}) = N_{\text{cases , subtype 2}} / N_{\text{poprisk}}$
  - Lum HER2: $\lambda_3(\text{age, year}) = N_{\text{cases , subtype 3}} / N_{\text{poprisk}}$
  - HER2 pos: $\lambda_4(\text{age, year}) = N_{\text{cases , subtype 4}} / N_{\text{poprisk}}$
  - TNBC: $\lambda_5(\text{age, year}) = N_{\text{cases , subtype 5}} / N_{\text{poprisk}}$

Note: Same N_{poprisk} for all incidence rates.
Only the numerator differs.

- **Incidence rates** can be estimated with Poisson regression for rates:
  - $\ln(\lambda_i(\text{age, year})) = \alpha * X_{\text{age}} + \beta * X_{\text{year}}$
  - $\ln(d_i(\text{age, year})) = \alpha * X_{\text{age}} + \beta * X_{\text{year}} + \ln(N_{\text{poprisk}})$
  - We ignored the calendar time effect (as only 2008-2019): $\ln(d_i(\text{age})) = \alpha * X_{\text{age}} + \ln(N_{\text{poprisk}})$
  - For stability, we estimated rates per 2-year age groups, i.e. <30, 30-31, 32-33, ..., 90-91, 92-93, 94+.

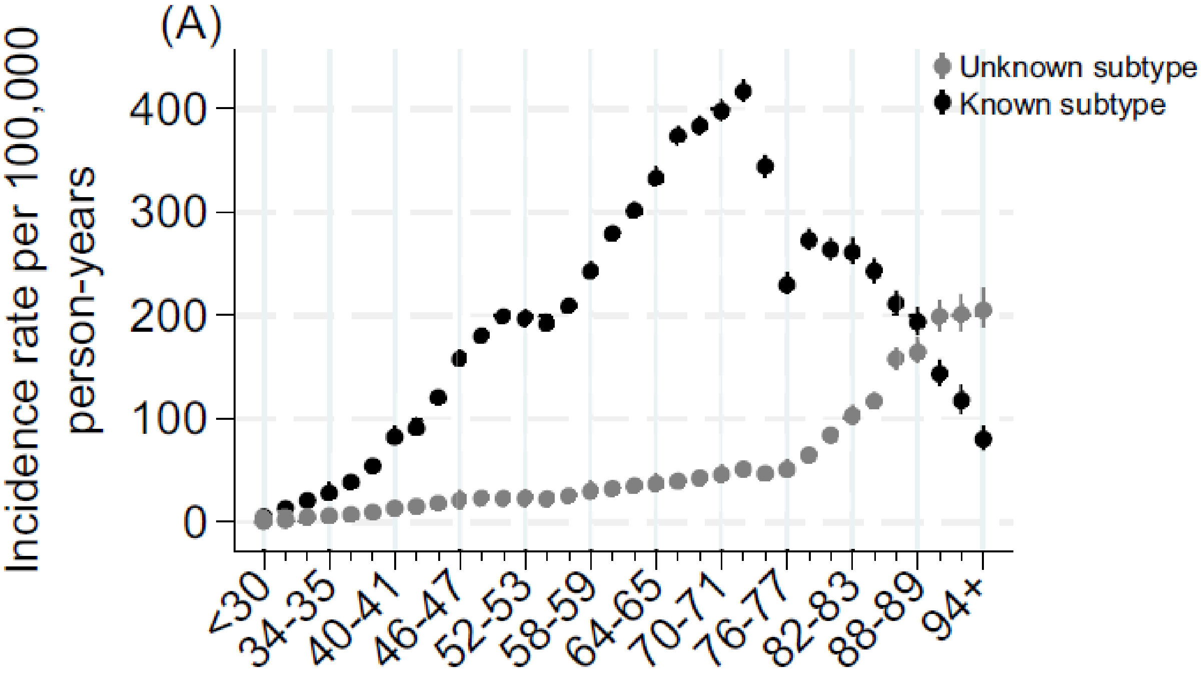
Missing data – on subtype

- Missing subtype: 15.5%; components had varying missing %
  - ER 4.1%, PR 4.2%, HER2 7.9%, grade 16.4%
  - **Leads to underestimation of the incidence rates with 15.5% - severe bias!**

- Reasons for missing
  - **Administrative missing** – not recorded in the earlier years, or in some regions/hospitals >> likely MCAR
  - Missing for **known reasons** – frail older women who are not eligible for certain treatments will not be investigated in detail, and thus will not be subtyped >> likely MAR (depending on known factors (age, treatment, comorbidity), which we can adjust for)
  - Missing for **unknown reasons** >> still likely MAR (depending on known factors)? Could also be MCAR/MNAR?

- Solution >> **Multiple Imputation with Chained Equations (MICE).**
  - Model-based method to impute missing data based on information in known variables (MAR).

Missing information depends on age



Multiple Imputation with Chained Equations (MICE)

- We have applied **multiple imputation with chained equations** (MICE).
- This involves a 2-step procedure
  - ✓ Step 1: Imputation model to generate *m* imputed (completed) datasets.
  - ✓ Step 2: Analysis (substantive) model (i.e. Poisson) to estimate incidence rates and pooling parameters across imputed datasets using Rubin’s rules #1 (mean) and #2 (variance).
- Key points:
  - Imputation model must capture the underlying reasons (patterns) for missing, e.g. age.
  - Imputation model must also reflect the analysis model, e.g. rates vary by age.
  - We added many clinical variables as auxiliary variables, and interactions with age (not trivial!). This will make the MAR assumption more likely.
- Implemented in Stata **mi package**:
  - Step 1: **mi impute**
  - Step 2: **mi estimate**

Building the imputation model

Variable	Missing of N=89,322 N (%)	Type in imputation	Imputation model (logistic, ordinal)	Analytical model (Poisson)
Age of diagnosis	Complete	Complete	X	X
Year of diagnosis	Complete	Auxiliary	X	
Region	Complete	Auxiliary	X	
Country of birth	Complete	Auxiliary	X	
Educational level	1,055 (1.18%)	Auxiliary / imputed	X	
Screening status	386 (0.43%)	Auxiliary / imputed	X	
T	251 (0.28%)	Auxiliary / imputed	X	
N	606 (0.68%)	Auxiliary / imputed	X	
M	Complete[a]	Auxiliary	X	
Stage (TNM)	465 (0.52%)		Not included (composite)	
Grade	14,610 (16.35%)	Imputed	X	
ER	3,657 (4.09%)	Imputed	X	
PR	3,778 (4.23%)	Imputed	X	
HER2	7,049 (7.89%)	Imputed	X	
Subtype	13,857 (15.51%)		Not included (composite)	X
Surgery	7,046 (7.89%)	Auxiliary / imputed	X	
Chemotherapy	17,818 (19.95%)	Auxiliary / imputed	X	
Radiotherapy	17,941 (20.08%)	Auxiliary / imputed	X	
Endocrine therapy	17,535 (19.63%)	Auxiliary / imputed	X	
Targeted therapy	17,992 (20.14%)	Auxiliary / imputed	X	
Survival indicators, CumHaz.(BC, OS)	Complete	Auxiliary	X	

Nelson-Aalen estimator of cumlative hazard of BC death and Overall death (OS)

[a]=deterministic prediction before imputation.

Step 1: Imputation model

- We used command **mi impute**.
- A total of *m*=30 imputed datasets, with 20 iterations, were generated.
  - *m* was chosen to be larger than the highest percentage of missing values among the variables (Targeted therapy=20.14%), according to the rule of thumb by White et al (2011).
- In the chained equations, each variable was regressed on all other variables
  - logistic regression models were applied to binary outcome variables (ER, PR, HER2, chemotherapy, radiotherapy, endocrine therapy, targeted therapy, screening status), while
  - ordinal regression was used for categorical outcome variables (grade, nodal involvement N, surgery type, T stage, educational level).
- The imputation models were **stratified by age at diagnosis** (using the ***by* option** in **mi estimate**)
  - thus corresponding to including **all pairwise interactions of age at diagnosis** (<40, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, ≥80) with **each** covariate in the model.
  - Important to include age-interactions! As our analysis model were modelling age-varying incidence rates.

Step 2: Analysis model and pooling

- We combined the (imputed) variables ER, PR, HER2 and grade into the **surrogate subtype variable** using the **mi passive:** command.
  - Using mi passive: we can create a new variable from the imputed variables.
  - Such a new variable can be used in mi commands.

- To each of the *m* imputed datasets, we fitted the **analytical Poisson model** with 2-year age groups as covariates; and saved the estimated parameters and standard errors.
  - This generated *m* sets of parameter estimates (and standard errors).

- To estimate **pooled parameters**, we manually applied Rubin's rules to the *m* parameters:
  - Rubin's rule #1: **Pooled parameter** is the mean of the estimated parameters for the imputed data sets (i.e. taking average of *m* parameters).
  - Rubin's rule #2: **Standard error of the pooled parameter**; incl. both within-between imputations variances.
- This is typically done by **mi estimate. I will come back to why we did not use mi estimate.**

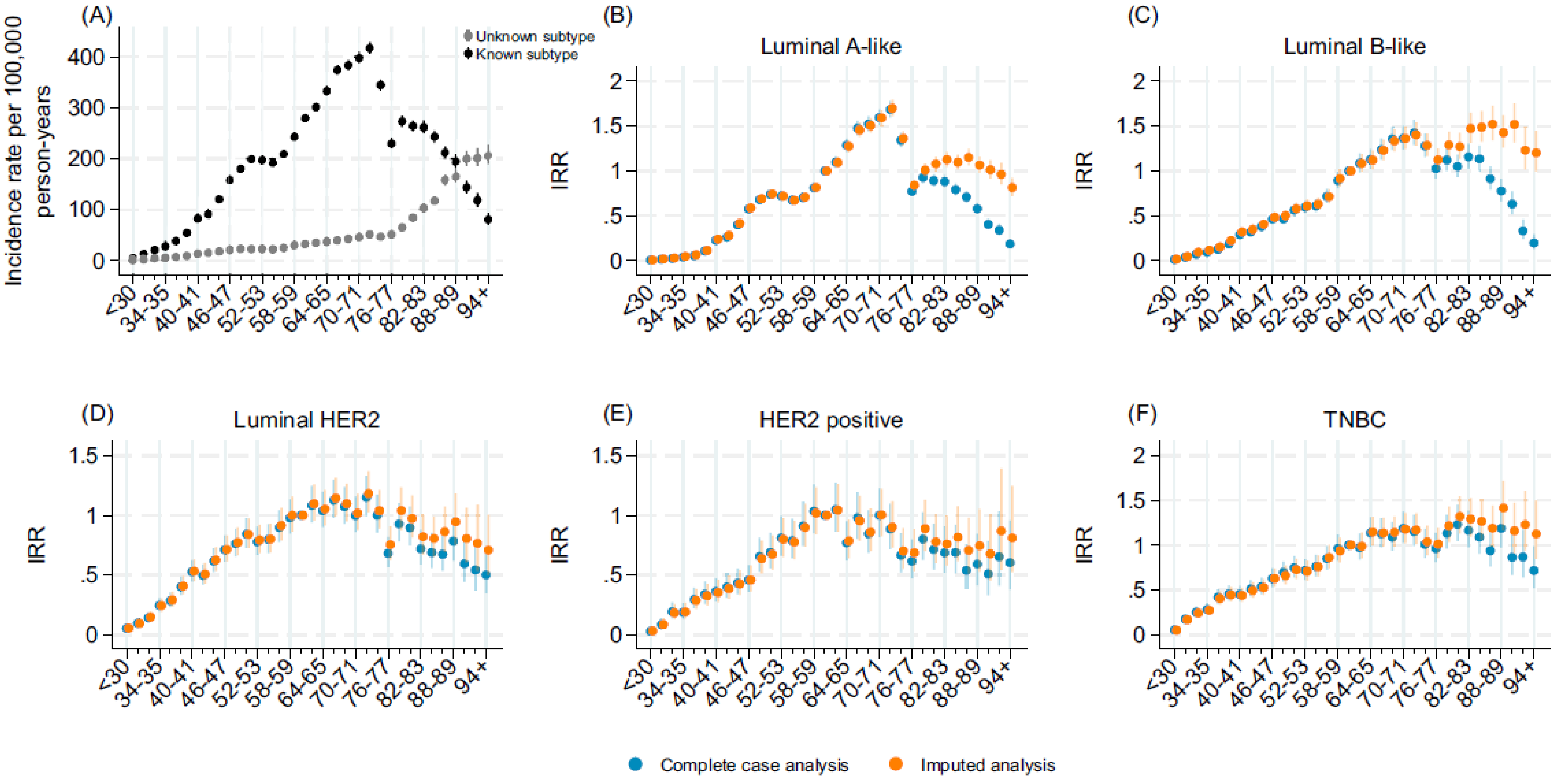


FIGURE 2 Breast cancer incidence rates of known and unknown subtype by age at diagnosis (A) and comparison of incidence rate ratios

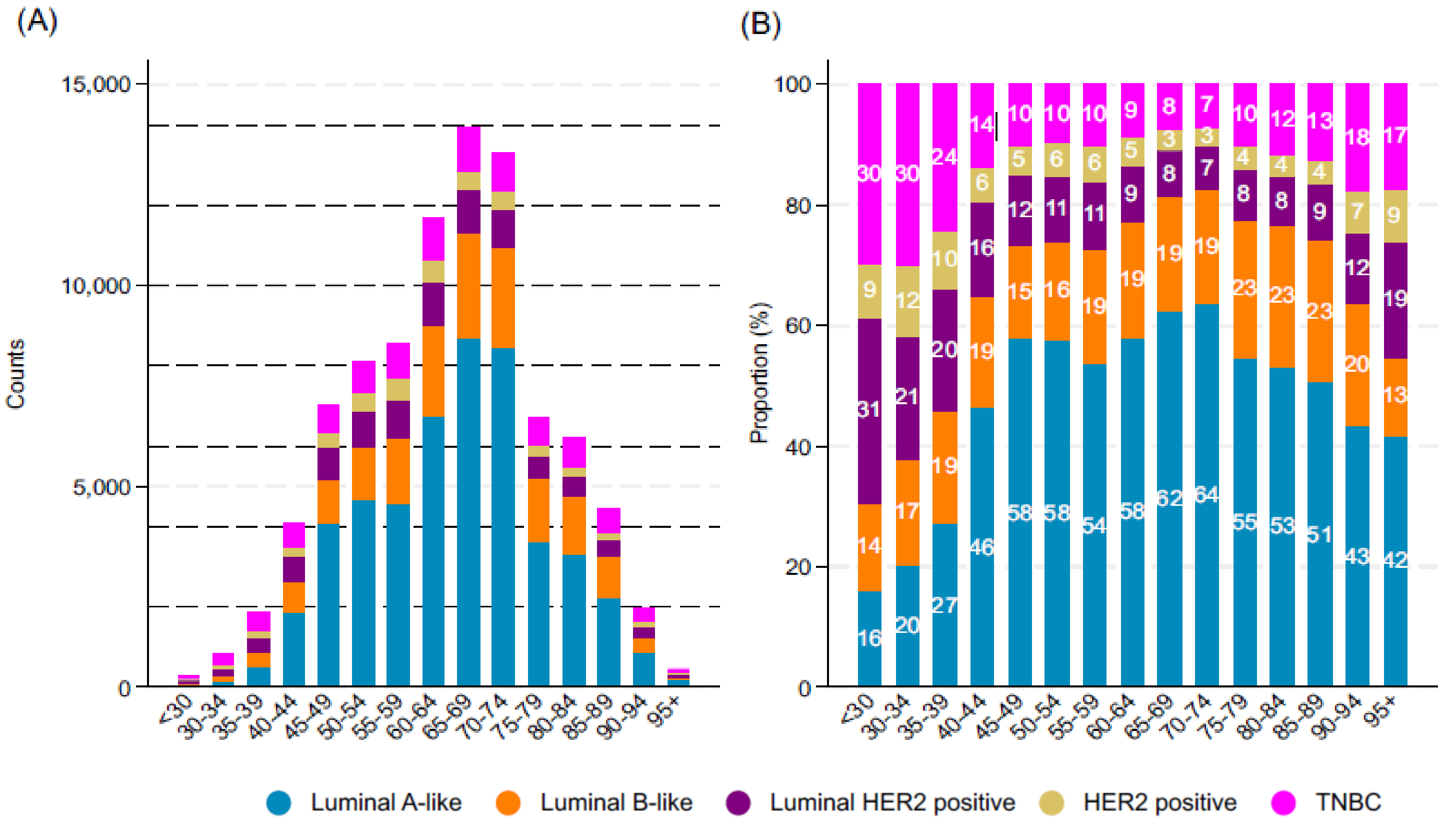
Appropriately estimating cancer incidence – without loss of cases

- A key strength of this analysis is that we included all women with missing subtype.
- Important as we were interested in the **absolute incidence rate** (not just relative rates):
  - If women with missing subtype are excluded - subtype-incidence rates will be underestimated.
- Relative effects are likely less impacted – at least in younger ages.

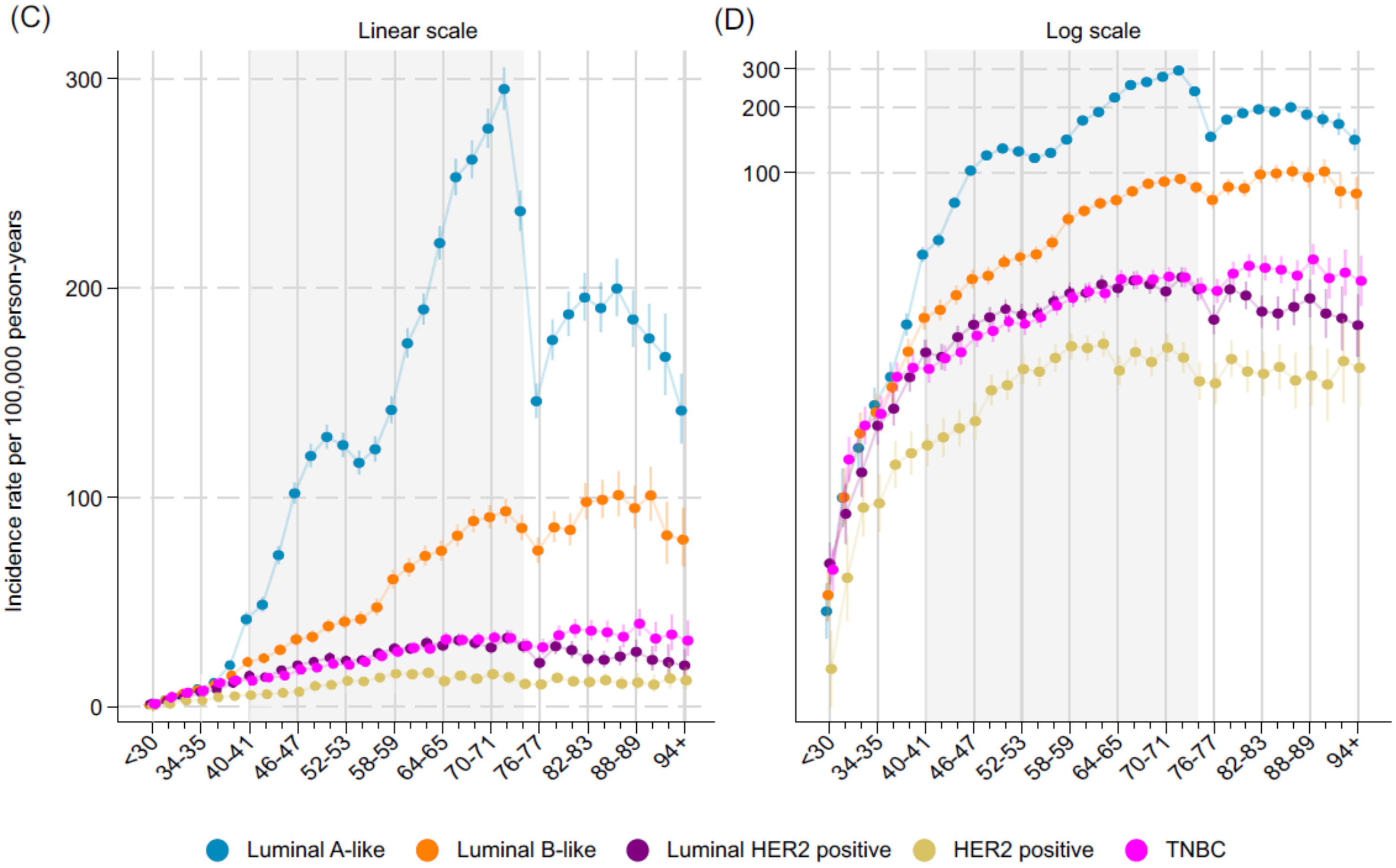
- The parameter estimates from the Poisson model represent
  - **Intercept**: rate in reference age
  - **Coefficients for covariates** (age effect): rate ratios (vs. the reference age)

- From these parameter estimates from the Poisson model, we can obtain at different ages
  - **Incidence rates** – directly from the parameters in the model.
  - **Proportions, counts** – by multiplying rate fractions of total rate (%) with the known N.
  - "pool-last-principle": to obtain **pooled** incidence rates & proportions:
    - transform rates/proportions from params in each imputed dataset >> then pooled them.

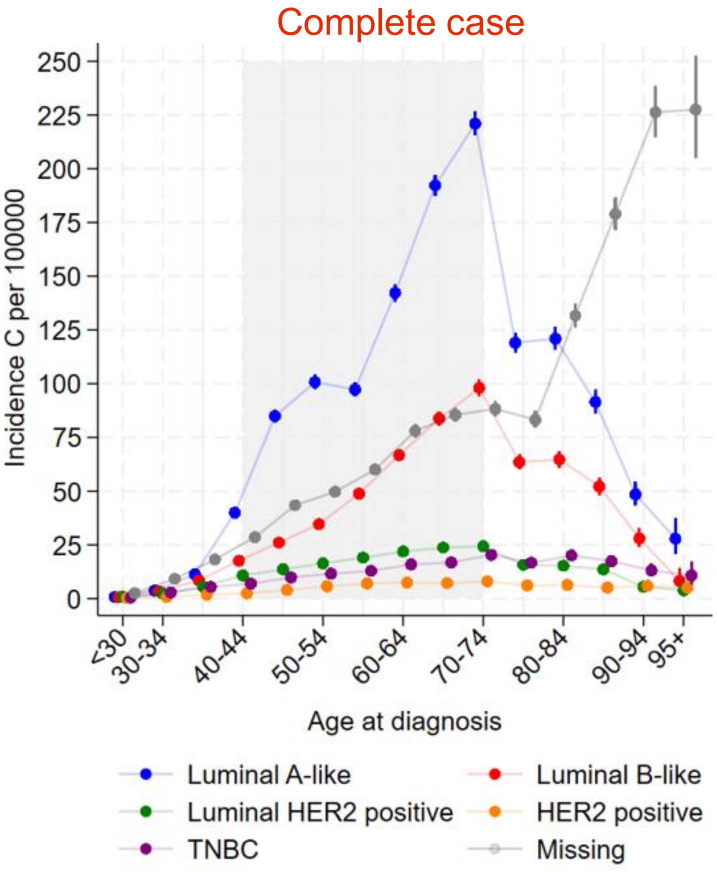
Results: Pooled counts and proportions post-estimated from parameters in Poisson model.

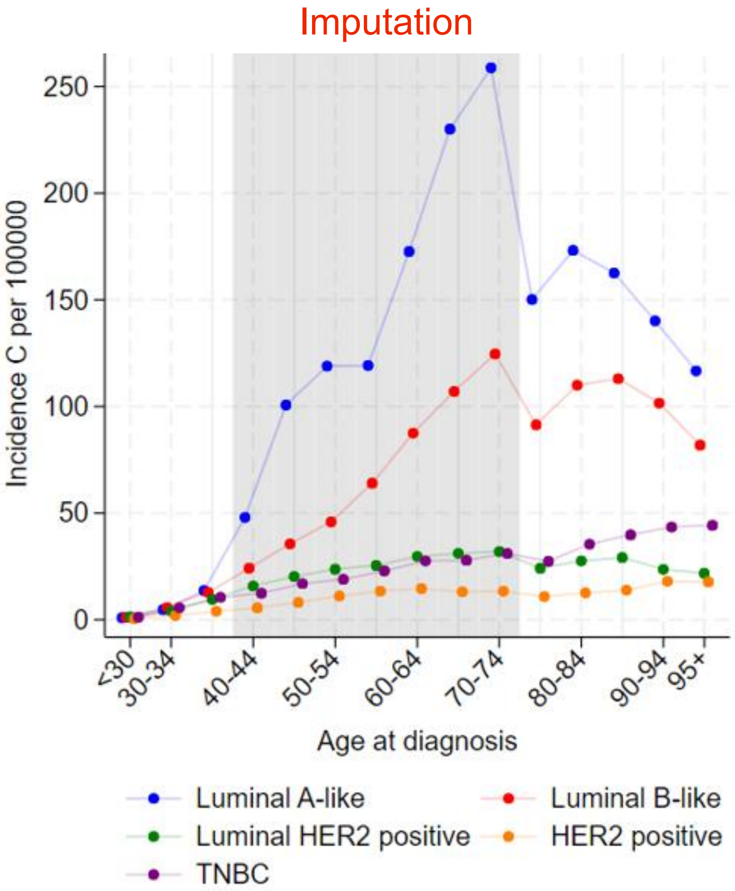


Results: Pooled incidence rates from parameters in Poisson model



Results not in the published paper! Absolute rates are higher in imputed analysis, esp. older ages.





Why we could not use **mi estimate** to obtain the pooled estimates

- MICE only works on individual-level data: We applied Poisson regression to individual level data.
- We only imputed the case dataset.
  - We then added the counts of "population-at-risk" for each individual, as an offset (in "pop_at_risk" variable).
  - E.g. **. poisson d_subtype1 i.age, exposure(pop_at_risk)**

- For each subtype: We restricted the case dataset to those with that subtype.
  - E.g. For the estimation of subtype 1 incidence – we only included the women with subtype 1.
- However, since we imputed the "outcome" (subtype):
  - Number of outcomes will differ for different *m* datasets.
    - Imputation m=1: 45000 Lum A
    - Imputation m=2: 44000 Lum A. etc.
  - This created **imputed datasets of different sizes for each subtype-specific model**.
  - As far as we could work out, **mi estimate** requires imputed datasets of the same size (same number of obs).
- In many applications of MICE, the interest is in imputing a covariate (exposure/treatment, confounder, etc.) – this will not change the size of the imputed datasets.

In conclusion

- This is one of the first studies to estimate subtype-specific BC incidence using MICE.
- MICE was important to avoid underestimation of incidence rates – up to 15.5% in our data.

- Essential to include a broad imputation model that captures the underlying missing data mechanism.
  - We had access to a wide range of clinical variables.
- Also important to include the correlation structures from the analysis model in the imputation model.

- **mi package** was a very powerful tool – it enabled us to estimate rates and impute the data.
  - Yet we faced challenges when using **mi estimate** for imputed datasets of different sizes.
- This is an example of **imputing the outcome** – while most examples in literature focus on imputing covariates (exposure/treatment, confounder) in regression models.

KAROLINSKA INSTITUTET * ANNO 1810 *

Karolinska
Institutet