



BMJ Open Danish Pathology Life Course (PATHOLIFE) cohort: a register based cohort extending upon a national tissue biobank

Pernille Yde Nielsen ¹, Andreas Bartholdy,¹ Lise Mette Rahbek Gjerdrum,^{2,3} Rudi Gerardus Johannes Westendorp ¹, Laust Hvas Mortensen,^{1,4} Samir Bhatt,¹ Majken Karoline Jensen¹

To cite: Nielsen PY, Bartholdy A, Gjerdrum LMR, *et al.* Danish Pathology Life Course (PATHOLIFE) cohort: a register based cohort extending upon a national tissue biobank. *BMJ Open* 2023;13:e068483. doi:10.1136/bmjopen-2022-068483

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2022-068483>).

Received 19 September 2022
Accepted 28 March 2023



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Department of Public Health, University of Copenhagen, Copenhagen, Denmark

²Department of Pathology, Zealand University Hospital, Roskilde, Denmark

³Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

⁴Data Science Lab, Statistics Denmark, Copenhagen, Denmark

Correspondence to

Dr Pernille Yde Nielsen;
pernille.yde@sund.ku.dk

ABSTRACT

Purpose The Danish Pathology Life Course (PATHOLIFE) cohort was established to facilitate epidemiological research relating histological and cytological features extracted from patient tissue specimens to the rich life course histories, including both prior and future register data, of the entire Danish population. Research results may increase quality of diagnosis, prognosis and stratification of patient subtypes, possibly identifying novel routes of treatment.

Participants All Danish residents from 1 January 1986 to 31 December 2019, totalling 8 593 421 individuals.

Findings to date We provide an overview of the subpopulation of Danish residents who have had a tissue specimen investigated within the Danish healthcare system, including both the primary sector and hospitals. We demonstrate heterogeneity in sociodemographic and prognostic factors between the general Danish population and the above mentioned subpopulation, and also between the general Danish population and subpopulations of patients with tissue specimens from selected anatomical sites. Results demonstrate the potential of the PATHOLIFE cohort for integrating many different factors into identification and selection of the most valuable tissue blocks for studies of specific diseases and their progression. Broadly, we find that living with a partner, having higher education and income associates with having a biopsy overall. However, this association varies across different tissue and patient types, which also display differences in time-to-death and causes of death.

Future plans The PATHOLIFE cohort may be used to study specified patient groups and link health related events from several national health registries, and to sample patient groups, for which stored tissue specimens are available for further research investigations. The PATHOLIFE cohort thereby provides a unique opportunity to prospectively follow people that were characterised and sampled in the past.

INTRODUCTION

Many Danish health registries date back decades and enable population-wide studies of patient groups by linking data for different health related events through the Civil

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ The Danish Pathology Life Course (PATHOLIFE) cohort includes histopathological information from archived tissue and cytology specimens covering >3million individuals from the Danish population, enabling large-scale epidemiological studies and relating histopathological features to relevant patient exposures and outcomes.
- ⇒ The PATHOLIFE cohort includes socioeconomic and health-related information from the entire Danish population from 1 January 1986 to 31 December 2019, enabling identification of clinically interesting patient groups and thorough investigation of possible selection bias, hence providing valuable external validation for research results.
- ⇒ Incomplete clinical information behind patient referral for biopsy and health related information from primary care and para-clinical exams and investigations, may increase potential bias in selected study populations.

Personal Registration (CPR) number, a unique identification number which is given to all Danish residents at birth or on immigration.^{1 2}

The Danish Pathology Life Course (PATHOLIFE) cohort includes full information from the Danish National Pathology Register (DNPR), which stores reports on cytological and histological specimens, collected within the Danish national healthcare system, including both the primary sector and hospitals.³ Furthermore, the PATHOLIFE cohort links individual data from several Danish registries, including information on hospitalisations, performed procedures and operations, diagnoses, use of prescription drugs, various socioeconomic information and information on causes of death. The PATHOLIFE cohort thus enables the construction of an individual patients' life course trajectory

with detailed information about health related events, see [figure 1](#).

Histological materials are of special interest for research purposes as these are in principle stored indefinitely in formalin fixed paraffin embedded blocks (FFPE-blocks), that are stored at room temperature in archives of Danish pathology departments. FFPE-blocks may be retrieved for extraction of additional information as new technologies become available.⁴ The PATHOLIFE cohort contains the necessary information for identification and localisation of these materials as well as the individual patients' health related events both before and after the material requisition date, thus making it possible to construct study populations of specific diseases or tissue types and relate to many possible exposures and outcomes.

Study populations based on select patient groups with archived FFPE-blocks constitute subcohorts within the PATHOLIFE cohort, that includes the entire Danish population. Such subcohorts are subject to selection bias as they present only the patients that were both offered and who accepted certain procedures. The PATHOLIFE cohort enables comparisons of subcohorts to the overall Danish population (and across different subcohorts), by quantifying demographic and socioeconomic differences, providing valuable information for external validation of research results. We showcase this and compare several subcohorts of patients that appear in the DNPR. Specifically, we compare the general Danish population to the 'DNPR subcohort' (all individuals that are registered within the DNPR), the 'FFPE-block subcohort' (all individuals with archived FFPE-blocks) and two examples of topography-specific subcohorts, namely, a 'Skin subcohort' and a 'Liver subcohort'.

COHORT DESCRIPTION

Population

The PATHOLIFE cohort includes the entire Danish population from 1 January 1986 to 31 December 2019. The PATHOLIFE cohort is thus equivalent to the general Danish population within this time period. This time period was chosen because demographic data, including dates of births, deaths, immigrations and emigrations, are available through Statistics Denmark. In the future, the PATHOLIFE cohort may be extended to cover a longer time period as updated register data becomes available. The PATHOLIFE cohort is purely register based and does not require informed consent or active involvement by the population. Within the above-mentioned time period, the Danish population increased gradually from 5.2 million persons to 5.9 million persons, see [figure 2A](#). Individuals enter the cohort on birth or immigration, and leave the cohort on death or emigration. Re-entry is possible but individuals are only under observation, during periods of residency in Denmark.

The 'DNPR subcohort' (defined in the introduction) includes a large part (58%) of the PATHOLIFE cohort, reflecting the fact that a large part of the Danish population

have had a tissue specimen investigated at least once during their life, see [table 1](#) and [figure 2](#). This is either due to pathology investigations performed in relation to illness (or suspected illness) or due to attendance in one or more national screening programmes.⁵ There are currently three national screening programmes: cervical cancer, breast cancer and colon cancer, leading to many occurrences of these anatomical sites in the DNPR. Since two of the three national screening programmes regard only women, there is a natural majority of women in the DNPR, see [table 1](#) and online supplemental figure 1. Some topographies (anatomical sites) are much more abundant in the DNPR than others and abundance also differs by year. We stress that many tissue specimens present in the DNPR are requisitioned due to referral from individual clinicians, and that the large abundance of the topographies related to national screening programmes (cervix, breast and colon) is only part of the total amount of tissue specimens registered in the DNPR. See the supplemental material for details on abundance of different anatomical sites over time (online supplemental table 1, figures 2 and 3).

The PATHOLIFE cohort uniquely identifies individual 'specimens' as materials from a unique requisition and of a unique topography. In order to identify, as accurate as possible, which specimens are archived in FFPE-blocks, the PATHOLIFE cohort combines information from Systematized Nomenclature of Medicine (SNOMED)⁶ codes and requisition-level variables in order to distinguish between histological and cytological specimens. See supplemental material for details on register variables that are used for identifying FFPE-blocks (online supplemental tables 2 and 3).

The female majority is less pronounced in the 'FFPE-block subcohort' as compared with the 'DNPR subcohort', see [table 1](#). The primary cause for this difference between the 'DNPR subcohort' and the 'FFPE-block subcohort' is a relatively large amount of cytological cervix specimens, requisitioned in relation to the national screening programme and registered in the DNPR, are not stored in FFPE-blocks (redundant cytological specimens are often stored for a few months before abolition).

The fraction of the general Danish population that appear in the 'DNPR subcohort' and the 'FFPE-block subcohort' increases over time and by 2019 these fractions are 60% and 52%, respectively, see [figure 2A](#). The number of individuals, that have specimens investigated, each year is shown in [figure 2B](#) (which includes all kinds of specimens) and in [figure 2C](#) (which includes only histological specimens for which FFPE-blocks are stored).

Available variables and data sources

All information in the PATHOLIFE cohort is collected from Danish national registries and linked through the CPR number. Information on tissue specimens is collected from the DNPR and is given in terms of SNOMED codes,⁶ which are alpha-numerical codes that translate into specifications of topography (T-codes), morphology

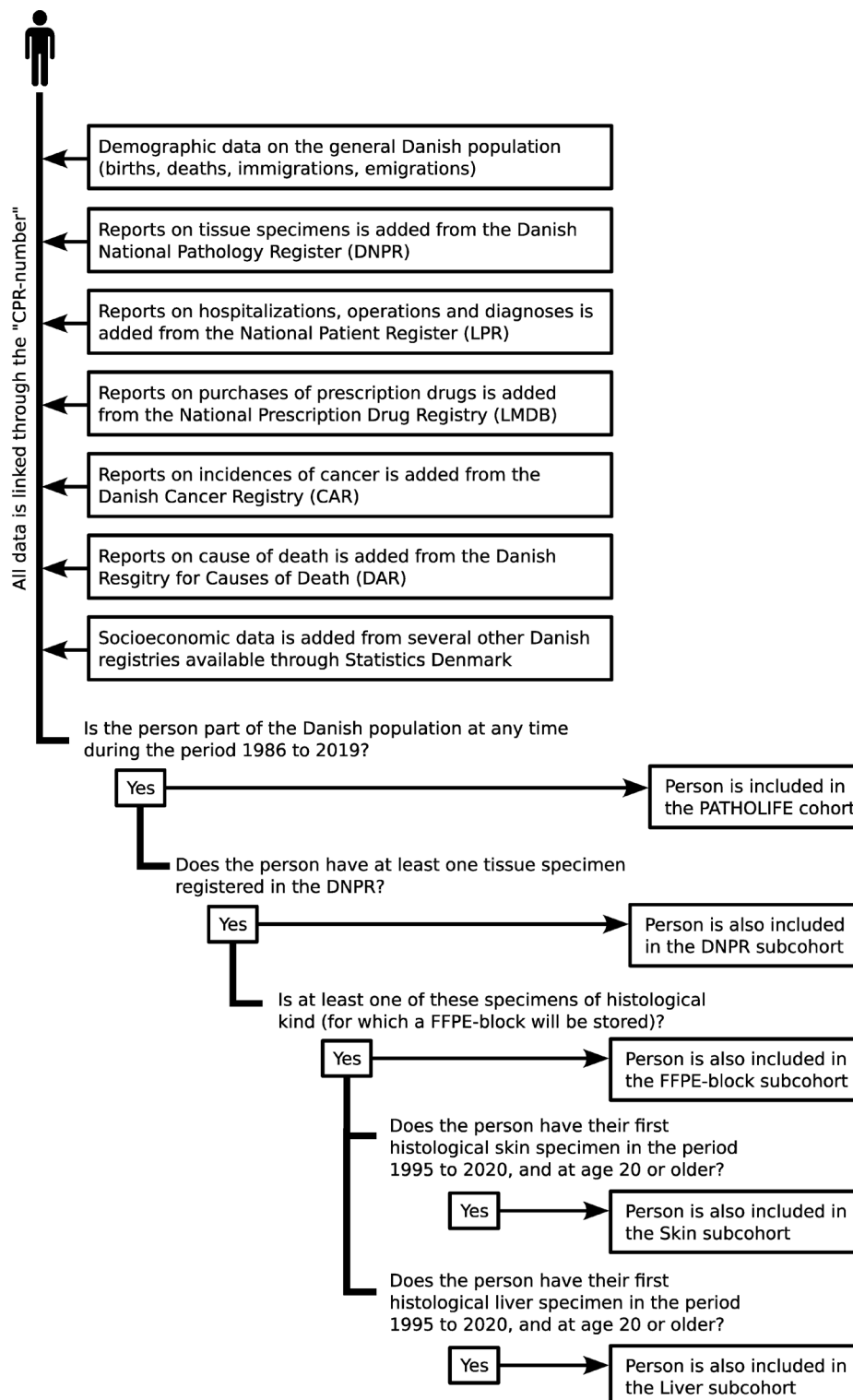


Figure 1 Flowchart showing how the different data sources are linked to the PATHOLIFE cohort and how subcohorts mentioned in the paper are defined. The PATHOLIFE cohort consists of all Danish residents from 1 January 1986 to 31 December 2019. Individual data sources are linked through the CPR number (individual identifier). Individual data points may date back before 1986. Some data sources may not have full coverage until a timepoint later than 1986 (see ‘Available variables and data sources’). The ‘DNPR subcohort’ consists of all individuals who have ever had a tissue specimen investigated and registered in the DNPR. The ‘FFPE-block subcohort’ is a subset of the ‘DNPR subcohort’ and consists of all individuals who have ever had a histological tissue specimen investigated and registered in the DNPR. All histological tissue specimens will have an FFPE-block stored indefinitely at a pathology department in a Danish hospital (see main text). The ‘Skin subcohort’ and the ‘Liver subcohort’ are both subsets of the ‘FFPE-block subcohort’. The ‘Skin subcohort’ and the ‘Liver subcohort’ may have some overlap, as it is possible to have both kinds of tissue specimens investigated. CPR, Civil Personal Registration; DNPR, Danish National Pathology Register; FFPE, formalin fixed paraffin embedded; PATHOLIFE, Danish Pathology Life Course.

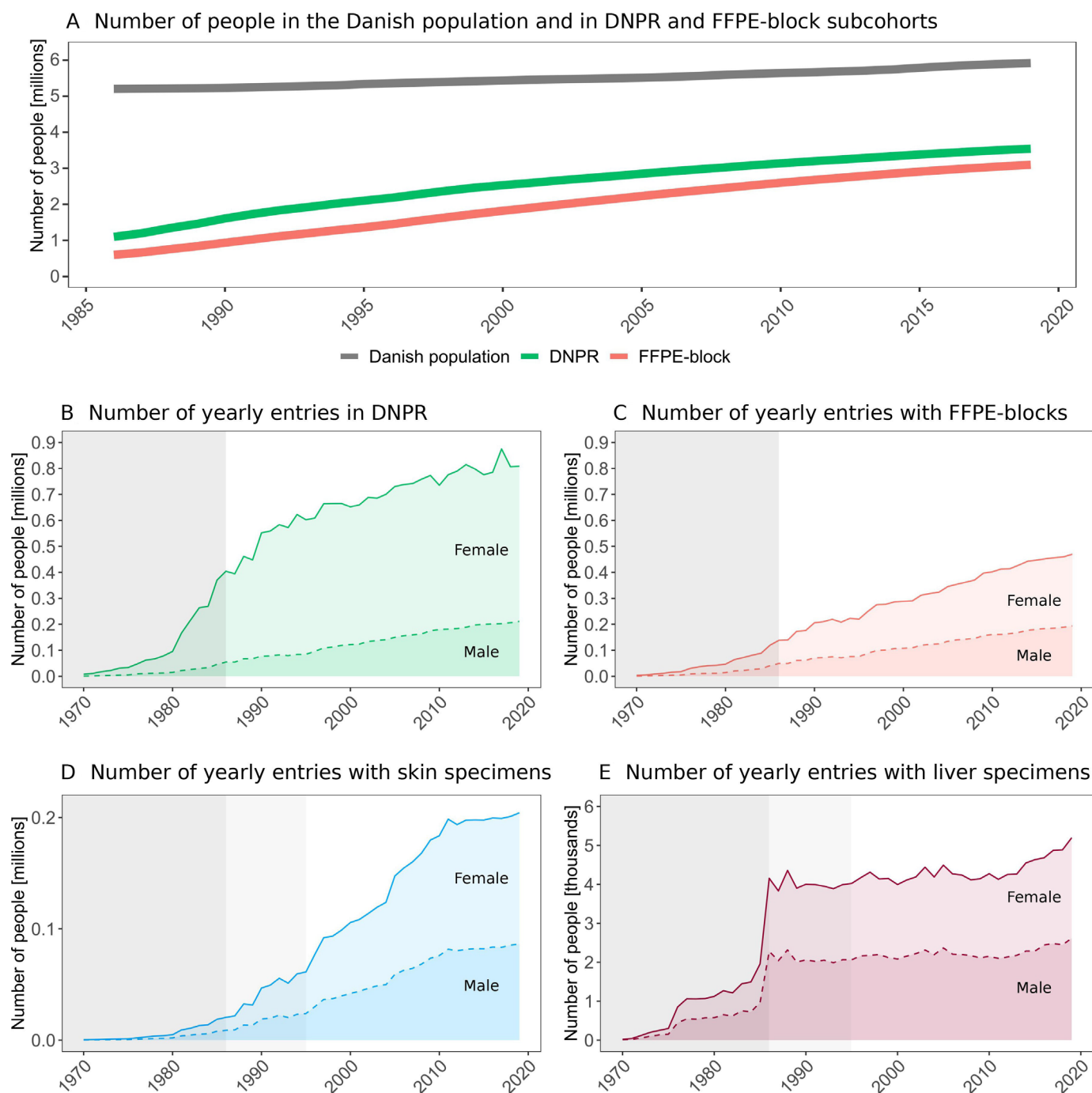


Figure 2 (A) The number of persons in the Danish population (grey), the number of persons which have at least one specimen in the DNPR (green) and the number of persons that have at least one FFPE-block in the DNPR (red). The numbers represent for each year the number of individuals who are alive at 1 January. For panel (A), individuals included in the 'DNPR subcohort' or the 'FFPE-block subcohort' may have had at least one tissue specimen during their life up until the given year. (B) The number of people that have specimens of any kind in the DNPR, counted each year. (C) The number of people that have histological specimens (and, therefore, also FFPE-blocks) in the DNPR, counted each year. (D) The number of individuals that have skin specimens in FFPE-blocks (topography code T01 or T02), counted each year. (E) The number of individuals that have liver specimens in FFPE-blocks (topography code T56), counted each year. In panels (B)–(E), individual persons may be counted once per calendar year, meaning that the same person may enter more than once, if the person has had several tissue specimens in different calendar years. The grey area displayed in panels (B)–(E) show data on tissue specimens that were investigated before 1986, when inclusion criteria for the PATHOLIFE cohort starts. The 'Skin subcohort' and the 'Liver subcohort' are restricted to include only individuals who have had their first histological specimen (of either skin or liver tissue) after 1 January 1995 and at an age of at least 20 years. This means that not all skin or liver specimens in the DNPR will enter the skin or liver subcohorts. The lighter grey areas displayed only in panel (D)–(E) shows data on tissue specimens that were investigated before 1995. DNPR, Danish National Pathology Register; FFPE, formalin fixed paraffin embedded; PATHOLIFE, Danish Pathology Life Course.

Table 1 Number of individuals included in the PATHOLIFE cohort and in the embedded subcohorts. For each column the number (and percentage) of M, F and NA is also reported

| | PATHOLIFE cohort | DNPR subcohort | FFPE-block subcohort | Skin subcohort | Liver subcohort |
|---------------------------|------------------|-----------------|----------------------|-----------------|-----------------|
| Number of individuals (N) | | | | | |
| Total | 8 593 421 | 4 951 708 | 4 333 258 | 1 922 965 | 82 128 |
| M | 4 222 264 (49%) | 2 013 351 (41%) | 1 926 489 (44%) | 840 707 (44%) | 41 485 (51%) |
| F | 4 213 007 (49%) | 2 932 862 (59%) | 2 403 023 (55%) | 1 082 258 (56%) | 40 643 (49%) |
| NA | 158 150 (2%) | 5495 (<1%) | 3746 (<1%) | – | – |

DNPR, Danish National Pathology Register; F, females; FFPE, formalin fixed paraffin embedded; M, males; NA, un-registered sex; PATHOLIFE, Danish Pathology Life Course.

(M-codes), procedures (P-codes), aetiology (Æ-codes) and diseases (S-codes). Also included from the DNPR are several administrative data and information in free text, describing the macroscopic and microscopic details and in some cases also results from molecular analyses. Information from the DNPR dates back to 1970, and has full national coverage from 1997.

Information on hospitalisations, operations and procedures performed at Danish hospitals, as well as diagnoses given to the patients in relation to hospital visits (ICD codes and SKS codes) are collected from the National Patient Register (LPR), and dates back to 1976.⁷

Information on purchases of prescription drugs (ATC codes) is collected from the National Prescription Drug Registry (LMDB), including data dating back to 1994. From 1997, LMDB also includes information on drugs given at Danish hospitals. Use of prescription drugs can be used to infer diagnoses that are not present in the LPR data.⁷

The PATHOLIFE cohort also includes variables from the Danish Cancer Registry (CAR) and the Danish Registry for Causes of Death (DAR).⁷

Information regarding age, sex, employment status, income, educational attainment, marital status, household characteristics, link to mothers' and fathers' CPR number and country of origin is also linked to PATHOLIFE, through registries provided by Statistics Denmark (www.dst.dk).⁸

Patient and public involvement

Patients and/or the public were not actively involved in the construction of the PATHOLIFE cohort.

FINDINGS TO DATE

Studies of clinical interest will often relate to specific diseases or diagnoses. However, study populations consisting of persons that have specific diagnoses or specific tissue specimens archived in FFPE-blocks, may include a highly heterogeneous mix of patient types. In order to bring attention to the existing heterogeneities and the possible confounding these heterogeneities may cause for researchers working with the archives samples, we show that subcohorts consisting of patients with FFPE-blocks of specific topographies may exhibit differences in

age (at time of biopsy), sex, prognosis (time to death and cause of death) and socioeconomic factors. We divided the 'Skin subcohort' and the 'Liver subcohort' into patient categories based on the findings of the biopsies and examined heterogeneity within same topography subcohorts.

We also showcase possible selection bias of the persons that may enter selected study populations, by assessing whether the persons that have tissue specimens investigated in the first place (ie, not regarding the findings in the biopsies) differ from the general population in terms of certain socioeconomic traits. We analysed the odds of having a certain type of specimen investigated within a certain year and computed the ORs for having such a specimen dependent on civil status, educational attainment and income level.

Skin and liver as example subcohorts

The number of individuals in the Danish population that have had skin biopsies taken, has increased with time and reached approximately 200 000 individuals by 2019, see [figure 2D](#) (where number of individual persons were counted per calendar year). The average fraction of females in the period 1970–2019 was 59%.

We constructed a 'Skin subcohort' consisting of individuals who have had their first skin specimen taken between 1995 and 2020. The subcohort was further restricted to include only skin specimens, that were archived in FFPE-blocks (ie, only histological specimens) and only persons of age ≥ 20 years at time of requisition—see details of subcohort definition in online supplemental material.

The 'Skin subcohort' was divided into five patient categories: Patients with malignant melanoma, patients with basal cell or squamous cell carcinoma, patients with other types of malignant cancer, patients with benign naevus and patients with other conditions. For details on how we define the different patient categories dependent on SNOMED codes from the DNPR, see online supplemental material. The five patient categories have different characteristic compositions of both sex, age at time of biopsy and time to death, see [table 2](#) and online supplemental figure 4.

We performed a similar investigation for liver. The number of individuals in the Danish population that have

Table 2 Characteristics of patient categories within the ‘Skin subcohort’

| | Malignant melanoma | Basal and squamous cell cancer | Other malignant cancer | Benign naevus | Other |
|---|--------------------|--------------------------------|------------------------|-------------------|-------------------|
| Number of individuals (N) | | | | | |
| M | 12 787 | 77 200 | 7232 | 161 756 | 581 732 |
| F | 14 111 | 74 239 | 10 664 | 309 009 | 674 235 |
| Age at requisition (average (IQR)) (years) | | | | | |
| M | 59.2 (48.6; 71.1) | 67.6 (59.1; 77.3) | 62.9 (53.6; 76.7) | 31.3 (19.5; 40.6) | 48.9 (34.6; 63.7) |
| F | 55.6 (41.6; 69.8) | 67.5 (57.2; 79.0) | 67.3 (57.7; 80.4) | 32.3 (21.5; 40.2) | 49.0 (33.9; 63.9) |
| Number of individuals who die before 2020 (N (%)) | | | | | |
| M | 3841 (30%) | 29221 (38%) | 3449 (48%) | 5556 (3%) | 94 254 (16%) |
| F | 3210 (23%) | 24 902 (34%) | 6231 (58%) | 7881 (3%) | 94 795 (14%) |
| Age at death (average (IQR)) (years) | | | | | |
| M | 74.2 (66.4; 84.0) | 81.9 (76.4; 88.7) | 76.0 (68.1; 86.0) | 65.2 (54.8; 78.0) | 75.8 (68.6; 85.0) |
| F | 78.0 (70.0; 88.8) | 85.4 (80.2; 92.4) | 77.2 (67.8; 88.3) | 66.9 (55.5; 80.1) | 79.4 (71.9; 89.1) |
| Follow-up time (average (IQR)) | | | | | |
| M | 7.93 (2.60; 11.9) | 7.96 (3.06; 11.8) | 6.78 (1.42; 19.4) | 11.9 (6.67; 17.3) | 9.97 (4.49; 14.7) |
| F | 9.46 (3.79; 14.3) | 8.67 (3.54; 12.8) | 6.33 (1.27; 9.77) | 12.7 (7.28; 18.4) | 10.5 (4.97; 15.4) |

F, females; IQR, Interquartile range; M, males.

had liver specimens taken has increased with time but stagnated around year 1985 and reached approximately 5000 individuals by year 2019, see [figure 2E](#) (where number of individual persons are counted per calendar year). The average fraction of females in the period 1970–2019 was 49%. The total number of persons with liver specimens is smaller than the corresponding number of persons with skin specimens, compare [figure 2D,E](#).

Similarly to the ‘Skin subcohort’, we constructed a ‘Liver subcohort’ consisting of individuals who have had

their first liver specimen taken between 1995 and 2020, restricted in the same way as the ‘Skin subcohort’.

The ‘Liver subcohort’ was divided into four patient categories: patients with malignant cancer, including metastasis, patients with hepatitis and cirrhosis related conditions, patients with no pathological findings and patients with other conditions. Although differences between patient categories in the ‘Liver subcohort’ were less pronounced than the differences observed in the ‘Skin subcohort’, the four patient categories did also

Table 3 Characteristics of patient categories within the ‘Liver subcohort’

| | Malignant cancer | Hepatitis/cirrhosis | No pathological findings | Other |
|---|-------------------|---------------------|--------------------------|-------------------|
| Number of individuals (N) | | | | |
| M | 20 376 | 12 251 | 2151 | 6707 |
| F | 17 761 | 12 162 | 2604 | 8116 |
| Age at requisition (average (IQR)) (years) | | | | |
| M | 67.9 (61.5; 75.5) | 51.6 (41.4; 63.3) | 59.6 (50.6; 71.5) | 58.0 (47.4; 71.4) |
| F | 67.4 (60.2; 76.0) | 54.4 (45.3; 65.6) | 58.5 (49.0; 70.2) | 56.1 (44.7; 69.8) |
| Number of individuals who die before 2020 (N (%)) | | | | |
| M | 18 347 (90%) | 5562 (45%) | 1384 (64%) | 3781 (56%) |
| F | 15 954 (90%) | 54.4 (40%) | 1444 (56%) | 3357 (41%) |
| Age at death (average (IQR)) (years) | | | | |
| M | 69.0 (62.6; 76.4) | 63.6 (55.9; 72.6) | 67.4 (60.0; 76.2) | 68.3 (61.4; 77.3) |
| F | 68.7 (61.6; 77.1) | 67.2 (58.7; 76.9) | 67.7 (59.6; 77.0) | 69.6 (62.1; 79.1) |
| Follow-up time (average (IQR)) | | | | |
| M | 1.09 (0.09; 1.05) | 8.02 (1.87; 13.2) | 5.65 (0.53; 8.28) | 6.24 (0.56; 10.6) |
| F | 1.16 (0.10; 1.19) | 8.31 (2.24; 13.6) | 6.96 (0.85; 12.1) | 7.78 (1.19; 13.3) |

F, females; IQR, Interquartile range; M, males.

exhibit some differences in characteristic compositions of both sex, age at time of biopsy and time to death, see [table 3](#) and online supplemental figure 5. Overall a larger fraction of the ‘Liver subcohort’ died before 2020 as compared with the corresponding fraction in the ‘Skin subcohort’ (compare [tables 2 and 3](#)).

Differences in top-ranking causes of death

We compared the highest ranking causes of death in the general Danish population and in the different subcohorts, for individuals who have died in Denmark in the period 1995–2019. While many top-ranking causes of death of the general population are also present as leading causes of death in the subcohorts, we also observe some differences, see [figure 3](#). The ‘Liver subcohort’ differed the most from of the general population, while the ‘Skin subcohort’ and many (but not all) of the embedded patient categories were in overall good agreement with the general population. The skin patient category with ‘benign naevus’, was a group we expected to be primarily biopsied due to suspicion of malignant melanoma. This group had top-ranking causes of death that were not very different from those of the general population. Markedly, the same is not true for the liver patient category with ‘no pathological findings’, which was a group we also expected to be primarily biopsied due to suspicion of malignant cancer or other disease. Here we found that all 10 top-ranking causes of death were malignant cancers. This difference may partly be explained by the liver being a relatively common metastatic site and also a tissue for which more invasive techniques are needed for biopsy, as compared with skin, for example. Regardless of the cause of this difference, it highlights how bias from indication may vary greatly between seemingly comparable ‘healthy’ tissue samples.

Examples of socioeconomic differences

Many factors may influence and bias which persons get biopsies from different anatomical sites. National screening programmes and the prevalence of illnesses cause bias in terms of specific topographies, and the individuals included in selected study groups.^{9 10} Indications and contraindications for undergoing the procedure(s) necessary for acquiring a tissue specimen may depend on the patient’s age, general health, lifestyle choices and other factors. Demographic and socioeconomic factors may influence the probability of having contact with the healthcare system and the inclination to accept or carry through the procedure.

In the following, we demonstrate some socioeconomic differences between the individuals that do and do not get (different kinds of) tissue specimens investigated (within the same year). Specifically, we looked at marital status, educational attainment and income—we noted that these specific factors represent only an arbitrary choice and not an exhaustive list of relevant factors to investigate. A series of logistic regressions were performed in order to estimate, for a given year, the OR (ie, odds of getting

a certain type of tissue specimen investigated and registered in the DNPR), depending on socioeconomic status (ie, living with a partner vs living alone; having completed elementary school vs not having completed elementary school; or having an income above median vs having an income below median). All analyses were stratified on sex, restricted to persons of age ≥ 20 years, and adjusted for categorical age (groups of 5-year intervals), see [figure 4](#).

Results revealed that, for the whole Danish population, living with a partner, having completed elementary school, and having an income above median was associated with a higher odds (OR >1) of getting a tissue specimen investigated within a given year. This is true for getting any kind of specimen, for getting a histological specimen and for getting a histological skin specimen (with few exceptions OR >1). However, the opposite was true for getting a histological liver specimen (with few exceptions OR <1). Statistical significance is expressed through the 95% CIs, which reveal that (in most cases) the ORs are significantly different from 1 (see [figure 4](#)). The exact values of 95% CIs are reported in online supplemental tables 4 and 5. We observed temporal trends in most of the investigated relationships, revealing that comparison of subcohorts over time may be influenced by—among many other things—changing compositions of socioeconomic factors. The overall trend was that ORs tended to increase with time (both for cases of OR >1 and OR <1).

Strengths and limitations

Through linking the DNPR with the entire Danish population alive on 1 January 1986, or born after this date, the PATHOLIFE cohort can be a useful tool in clinical epidemiological research on patients with histological and cytological features from tissues. The PATHOLIFE cohort includes many health related events and outcomes, which can facilitate identification and analysis of clinically interesting patient groups. Furthermore, the PATHOLIFE cohort enables identification of physically stored histological specimens and while we have shown, that the number of specimens is highly variable depending on, for example, anatomical site or patient age, in general sufficient samples exist for large scale surveys and analysis.

For studies aiming to compare progression patterns (ie, pathological features or biomarkers measured in biopsies obtained from patients) in patients that meet the same inclusion criteria, the potential bias that could come from conditioning study inclusion on having a tissue specimen should be considered. Disease burden has a complex relationship with socioeconomic and demographic factors leading to systematic biases between groups. Problematically, these systematic biases may change over time and are likely to be present at ever finer stratifications.¹¹

A key strength of the PATHOLIFE cohort is, therefore, its integration within the Danish national register system, including both health related, socioeconomic and demographic information, hence enabling thorough investigation of possible selection bias. This broad combination of variables may provide valuable external validation of

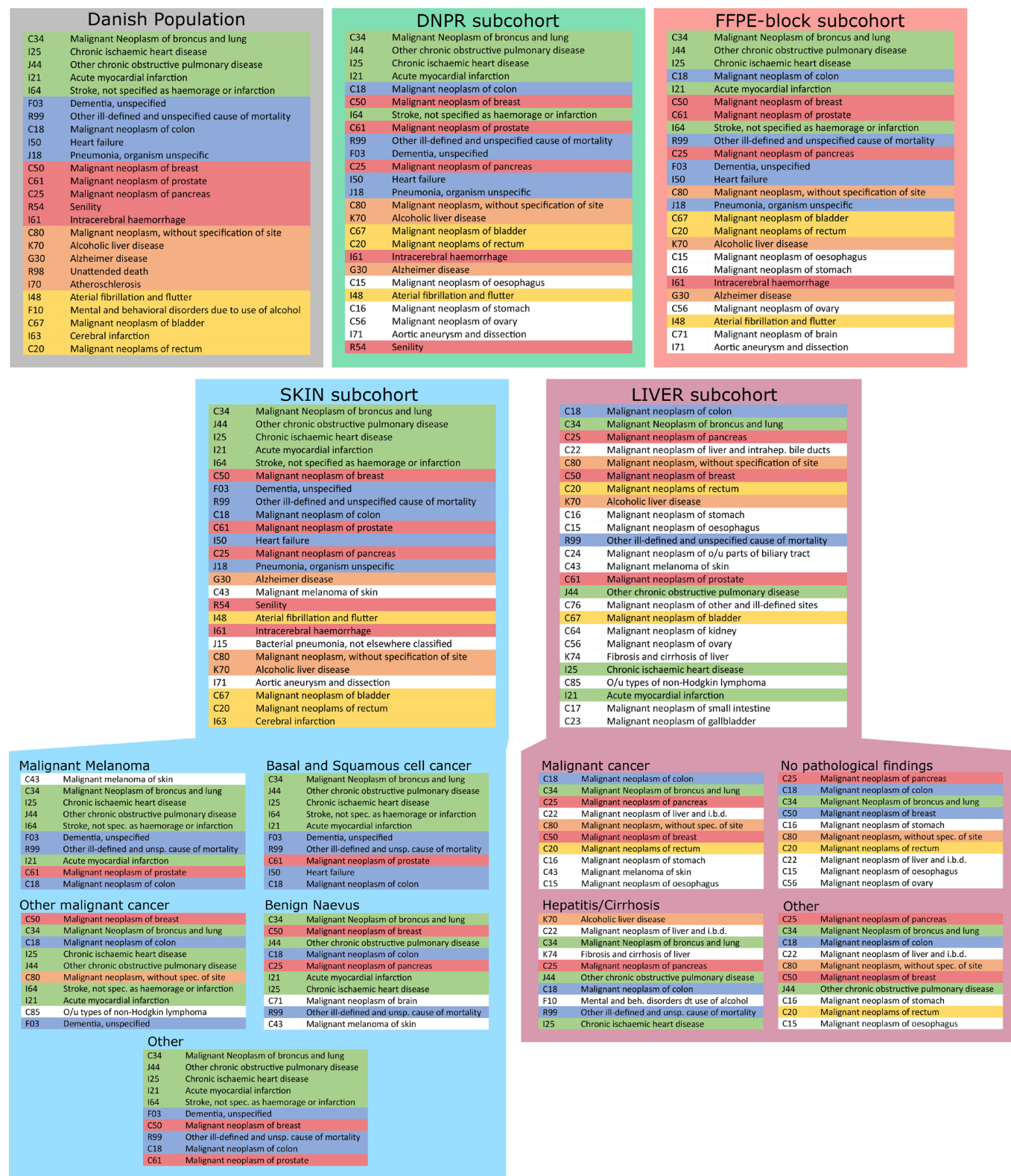


Figure 3 Top-ranking causes of death, for individuals who have died in Denmark in the period 1995–2019 (both years included). Top-ranking causes of death are listed for the entire Danish population, for individuals in the DNPR, for individuals with FFPE-blocks, for individuals in the ‘Skin subcohort’ and for individuals in the ‘Liver subcohort’. The individual causes of death are coloured according to rank in the entire Danish population (top 5 are green, top 6–10 are blue, top 11–15 are red, top 16–20 are orange and top 21–25 are yellow). Causes of death that are not coloured (white) are causes that do not appear in the top 25 causes of death in the entire Danish population but may appear in one or several subpopulations. The individuals with skin and liver specimens are further divided into patient subgroups as described in the main text, and the top 10 causes of death are shown for each subgroup, displaying heterogeneous compositions of top-ranking causes of death for the different patient subgroups. DNPR, Danish National Pathology Register; FFPE, formalin fixed paraffin embedded.

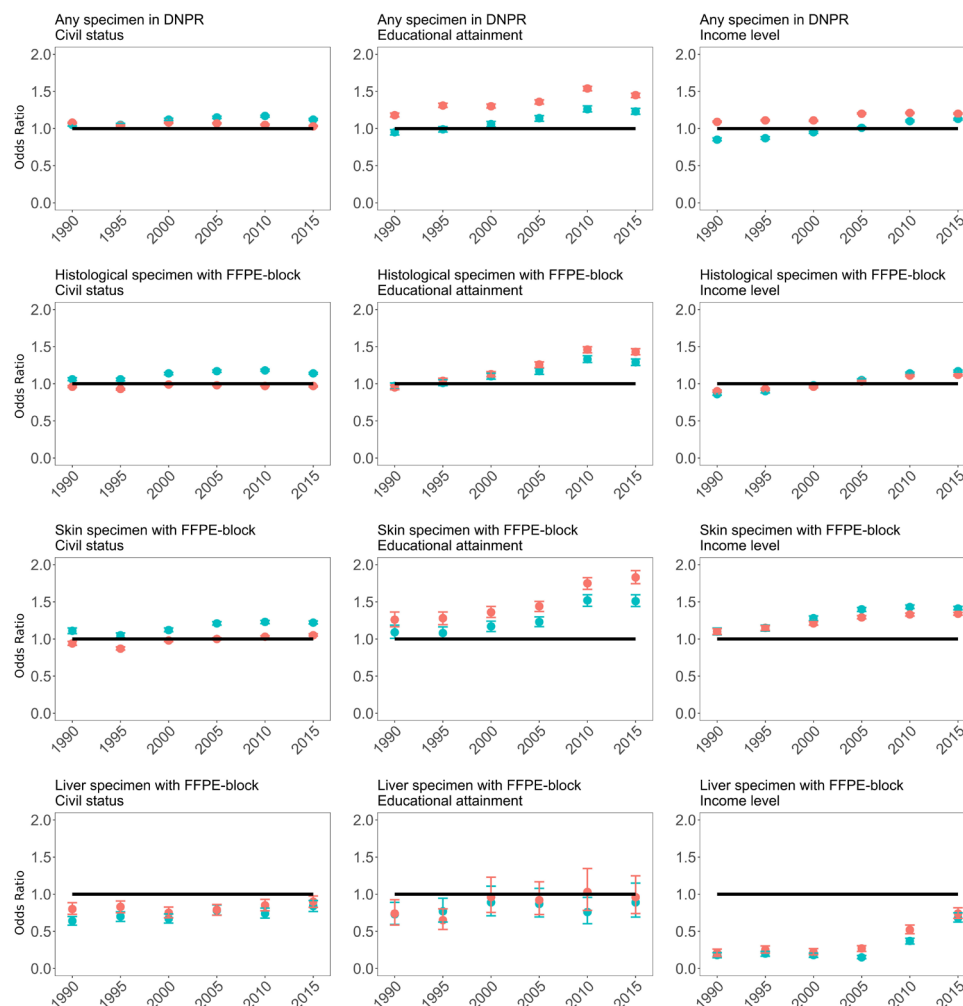


Figure 4 Logistic regressions were performed assessing the ORs for having at least one tissue specimen investigated within a given year. Top row displays ORs for having any kind of specimen. Second row displays ORs for having a histological specimen (and hence an FFPE-block, which is stored for all histological specimens). Third row displays ORs for having a histological skin specimen. Fourth row displays ORs for having a histological liver specimen. All analyses includes the entire Danish population alive at 1 January of the years indicated. Left column displays the OR for persons living with a partner versus persons living alone. Middle column displays the OR for persons that have versus have not completed elementary school. Right column displays the OR for persons with an income above versus below median. Black lines indicate an OR of 1. All analysis are restricted to include persons of age 20 years or older. Statistical significance is expressed through the 95% CIs, indicated by vertical bars (exact values of ORs and 95% CIs are reported in online supplemental tables 4 and 5). DNPR, Danish National Pathology Register; FFPE, formalin fixed paraffin embedded.

research results and patient groups can readily be stratified by, for example, socioeconomic and demographic features. There are, however, several limitations.

First, not all health related information is available, including patient information from primary care, para-clinical exams and observations (eg, bloodwork) or from bedside examinations (eg, body mass index or smoking status). A lack of primary care data is a substantial limitation given this is the entry point for early diagnoses and generally a point of care which individuals will access repeatedly until referral to secondary care. This limitation can be somewhat mitigated through other registers, for example, prescription drugs, but will inevitably remain a data gap.

Another limitation is the lack of clinical information surrounding the reason for referral. Indications for

retrieval of a tissue biopsy includes the diagnostic and prognostic value presumed to be obtained from such an investigation. Contraindication includes assessment of the individual risk imposed by the procedure and possibly also the economic cost. Both of these factors are expected to be dependent on the patients' overall health status and 'suspected illness severity'.

Overall health status and 'illness severity' is also expected to be associated with disease outcome, risk of comorbidity, etc. Such selection bias, where there is a (an unknown) common cause of both study inclusion and outcome, can distort the observed relationships observed between exposures and outcomes.

While there might be no easy solution to such a selection problem due to the unknown reason for referral, a large number of statistical techniques exist to 'correct'

these biases in the observational data, for example, inverse probability weighting, double machine learning and targeted maximum likelihood estimation,^{12 13} and the PATHOLIFE cohort includes many variables to enable the use of such methods.

COLLABORATION

Access to the Danish register data must be granted by Statistics Denmark and The Danish Health Data Authorities and is, therefore, not open access. Access to the established PATHOLIFE cohort is available to other investigators through collaborative agreements and a secured access. Please contact Professor Majken K Jensen (maje@sund.ku.dk) for further information.

Acknowledgements We acknowledge the researchers in the BIO-EPI Research Group at the Section of epidemiology, Department of Public Health, University of Copenhagen, who have been involved in setting up the Danish Pathology Life Course cohort and writing of data authorisation applications.

Contributors MKJ, LHM and RGJW initialised the construction of the Danish Pathology Life Course cohort. PYN conceptualised the paper, wrote the original draft and was responsible for data management and analysis. PYN, AB, LMRG, SB and MKJ contributed to the writing of the manuscript. PYN is responsible for the overall content as guarantor. All authors have critically revised and approved the final version.

Funding The study was funded by the Novo Nordisk Foundation Challenge Programme (NNF170C0027812).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Human participants are only passively involved through national register data, which does not require approval from ethics committee. Approval to use register data was granted by Statistics Denmark and the Danish Health Data Authority.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Access to the Danish Pathology Life Course cohort is available through collaborative agreements and granted access by Statistics Denmark and the Danish Health Data Authorities. Please contact Professor Majken K Jensen (maje@sund.ku.dk) for further information.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Pernille Yde Nielsen <http://orcid.org/0000-0003-4412-8084>

Rudi Gerardus Johannes Westendorp <http://orcid.org/0000-0002-0672-8372>

REFERENCES

- 1 Pedersen CB. The danish civil registration system. *Scand J Public Health* 2011;39:22–5.
- 2 Thygesen LC, Ersbøll AK. Danish population-based registers for public health and health-related welfare research: introduction to the supplement. *Scand J Public Health* 2011;39:8–10.
- 3 Erichsen R, Lash TL, Hamilton-Dutoit SJ, et al. Existing data sources for clinical epidemiology: the danish national pathology registry and data bank. *Clin Epidemiol* 2010;2:51–6.
- 4 Mund A, Coscia F, Kriston A, et al. Deep visual proteomics defines single-cell identity and heterogeneity. *Nat Biotechnol* 2022;40:1231–40.
- 5 Danish Health Authorities. 2022. Available: www.sst.dk/da/viden/screening/screening-for-kraeft
- 6 Patobank. 2022. Available: www.patobank.dk
- 7 The Danish Health Data Authority. 2022. Available: www.esundhed.dk/dokumentation
- 8 Statistics Denmark. 2022. Available: www.dst.dk/da/tilsalg/forskningsservice/data
- 9 Howe LD, Tilling K, Galobardes B, et al. Loss to follow-up in cohort studies: bias in estimates of socioeconomic inequalities. *Epidemiology* 2013;24:1–9.
- 10 Pallesen AVJ, Herrstedt J, Westendorp RGJ, et al. Differential effects of colorectal cancer screening across sociodemographic groups in Denmark: a register-based study. *Acta Oncologica* 2021;60:323–32.
- 11 Mierau JO, Turnovsky SJ. Demography, growth, and inequality. *Econ Theory* 2014;55:29–68.
- 12 Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econom J* 2018;21:C1–68.
- 13 van der Laan MJ, Rubin D. n.d. Targeted maximum likelihood learning. *Int J Biostat*;2.