

DESIGNING A CLINICAL STUDY FOR PANCREATIC CANCER RISK FACTORS

Annotated exercise sheet

1 Introduction

The purpose of this activity is to give students the opportunity to think about the issues involved in designing case-control and cohort studies (future development: think about adding a cross-sectional study and a randomised control trial).

2 Setting the scenario

Pancreatic cancer is the 10th most common cancer with 10,449 people diagnosed with pancreatic cancer in the UK in 2018. It has the lowest survival of all common cancers, with five-year survival less than 7%.

Early diagnosis is crucial to improve survival outcomes, with one-year survival in those diagnosed at an early stage six times higher than one-year survival in those diagnosed at stage four. However, most people with pancreatic cancer are diagnosed at a late stage (UK statistics on <https://www.pancreaticcancer.org.uk/>).

Patients usually experience no specific symptoms until the cancer is in a late stage. Currently there is no screening programme for pancreatic cancer in the UK, nor are there data to show reduced mortality from pancreatic cancer from any screening test. The diagnosis is usually made by imaging, after late symptoms such as jaundice and weight loss have already developed. Some data provide preliminary support for use of imaging methods (endoscopic ultrasound and MRI) if patients screened are **at increased risk** of pancreatic cancer. Therefore an accurate, non-invasive way to identify people at sufficiently high risk could enable the development of pancreatic cancer screening.

Known risk factors for pancreatic cancers include age, sex, smoking, high body mass index (BMI), diabetes, certain chemicals, ethnicity, and family history. There have also been investigations on the utility of a variety of **biomarkers** to identify those at sufficiently increased risk of pancreatic cancer to justify screening using imaging.

Cohort studies have a long history, but a recent development has been the use of biobanks from large population cohorts (such as UK Biobank - <https://www.ukbiobank.ac.uk>), that record data on participants from traditional questionnaires, in addition to biological samples, for instance to help evaluate the association between genetic variation, environmental exposures for risk of disease. Disease-specific biobanks have also been developed. One such at Barts Cancer Institute is the Pancreatic Cancer Research Fund Tissue Bank (PCRF, for further details see <https://www.thepancreastissuebank.org>). Set up in 2016, it has collected:

- Blood samples from 2,200 consenting patients who underwent biopsies or surgery for pancreatic diseases, including pancreatic cancers (also urine, saliva and tissue samples);
- Large numbers of healthy control blood (also urine and saliva).

2.1 Open questions

Question 2.1. “How would you design a study to investigate whether or not a proposed biomarker (or a combined panel of biomarkers) is an effective tool for identifying those at high risk of pancreatic cancer, in order to enable early detection of pancreatic cancer?”

- **Cross-sectional study:** take biological samples, evaluate the biomarker (or panel) at the same point in time as a diagnostic test for pancreatic cancer. For instance, run the study in those attending a GP appointment, or better, target a high risk group – eg. If registry people with strong family history for pancreatic cancer, or patients with diabetes mellitus. Would be difficult to run - no long term outcome; and pancreatic cancer is very rare so study would have to be enormous to find even a handful of cases.
- Could do a **case-control** study to assess the performance of the biomarker in detecting pancreatic cancers across all stages. Any of the following might be discussed (don't need to go through them all if not though)
 - Collect samples from patients with a new diagnosis of pancreatic cancer (blood and tissue) and select healthy controls without a diagnosis of cancer (blood). Think about who controls will be recruited and how.
 - Use a biobank focused on pancreatic cancer for selecting participants with cancer, and controls without cancer from the same biobank (e.g. the Pancreatic Cancer Research Fund Tissue mentioned above <https://www.thepancreastissuebank.org>; controls might have benign). Or use a cohort or trial from a population without pancreatic cancer at entry (e.g. UK Biobank <https://www.ukbiobank.ac.uk>) where healthy patients are recruited and followed up, with biological samples stored in a biobank. Here one might also do a **nested case-control** study by taking cases diagnosed within a certain time of entry, and matched controls, and use the biobank samples to run the tests. Suggest that we are working under the assumption that technology works on the available material and that storage does not degrade samples.
- Do a **cohort** study to validate the use of the test in detecting pancreatic cancers earlier. Remark that the number of participants would be huge for a general population sample (eg. young), so likely need to enrich by recruiting a higher-risk group using other factors (e.g. diabetes mellitus, strong family history, ...). Think about possible ethical issues.
- Could do a **randomised controlled trial**, where we use the biomarker to take some sort of action (screening with imaging, clinical work-up), but also here one must be careful of ethical issues (e.g. participants whose tests detects potential signals of cancer, or identified at very high risk, will need to be told their group and be referred for further tests. Would need evidence that the test does identify high risk before being used to take clinical actions).

Note 2.2. Be careful about identifying cohort and case-control studies as **prospective** and **retrospective** studies. Cohort and case-control are more accurate, especially as you can do a prospective case-control study (nested case-control – where at the study's end, the cases and controls are sampled from a cohort) and you can do a retrospective cohort study where you collect a cohort from historical data.

Question 2.3. “What summary measure of risk associated with the biomarker could you use for each study?”

Classical metrics:

- Cohort study: relative risk – for example risk of pancreatic cancer in people who were biomarker positive divided by risk in people who were biomarker negative, if binary. If continuous: relative risk per unit of the biomarker, or per SD of the biomarker. Also could consider a hazard or rate ratio, where the time-to-event is considered (where the event can be the diagnosis).
- Case-control study: odds ratio (again could be binary or continuous per unit etc).

Metrics for diagnostic tests (not covered in preliminary material):

- *Sensitivity: the proportion of test-positive subjects out of all disease-positive subjects. Independent of prevalence.*
- *Specificity: proportion of test-negative subjects out of all disease-negative subjects. Independent of prevalence.*
- *Positive Predictive Values (PPV): probability that following a positive test result, that individual will truly have that specific disease. Depends on prevalence.*
- *Negative Predictive Values (NPV): probability that following a negative test result, that individual will truly not have that specific disease. Depends on prevalence.*

Summary: *either take people without disease, evaluate the biomarker and follow them up to assess risk in each group, hence RR or HR; or take people with disease and without disease, and assess the biomarker (odds ratio). Remark that OR is equivalent to RR and HR if the disease is rare.*

3 Group activity

3.1 Work in small groups

“Assume that Cancer Research UK is going to fund two different studies to investigate whether your research group’s biomarker panel can help stratify risk of pancreatic cancer, and enable early detection that might lead to improvements in public health outcomes. One will be a case-control study, that will report its results within 3 years and the second group they need to design a cohort study which will report its results within 13 years.”

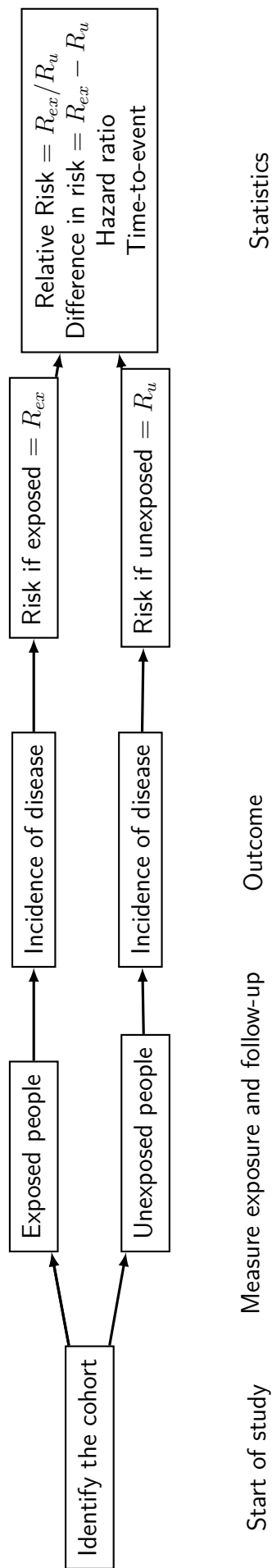
Group A (and multiples) needs to design a case-control study that reports results within 3 years;

Group B design a cohort study that reports results within 13 years.

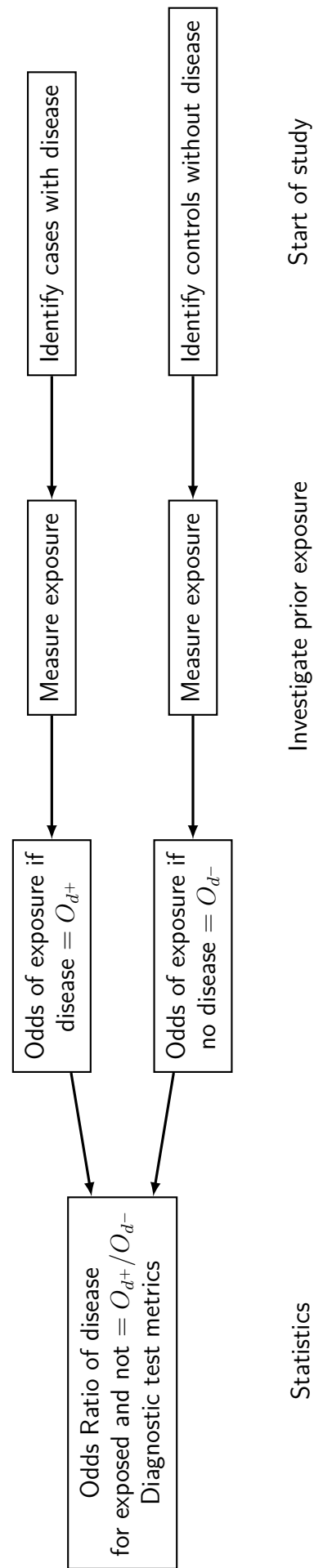
4 Annotated students' materials

4.1 Cohort and case control studies and possible metrics

Cohort study



Case-control study



Pancreatic cancer is the 10th most common cancer with 10,449 people diagnosed with pancreatic cancer in the UK in 2018. It has the lowest survival of all common cancers, with five-year survival less than 7%.

Early diagnosis is crucial to improve survival outcomes, with one-year survival in those diagnosed at an early stage six times higher than one-year survival in those diagnosed at stage four. However, most people with pancreatic cancer are diagnosed at a late stage (UK statistics on <https://www.pancreaticcancer.org.uk/>).

Patients usually experience no specific symptoms until the cancer is in a late stage. Currently there is no screening programme for pancreatic cancer in the UK, nor are there data to show reduced mortality from pancreatic cancer from any screening test. The diagnosis is usually made by imaging, after late symptoms such as jaundice and weight loss have already developed. Some data provide preliminary support for use of imaging methods (endoscopic ultrasound and MRI) if patients screened are **at increased risk** of pancreatic cancer. Therefore an accurate, non-invasive way to identify people at sufficiently high risk could enable the development of pancreatic cancer screening.

Known risk factors for pancreatic cancers include age, sex, smoking, high body mass index (BMI), diabetes, certain chemicals, ethnicity, and family history. There have also been investigations on the utility of a variety of **biomarkers** to identify those at sufficiently increased risk of pancreatic cancer to justify screening using imaging.

Cohort studies have a long history, but a recent development has been the use of biobanks from large population cohorts (such as UK Biobank - <https://www.ukbiobank.ac.uk>), that record data on participants from traditional questionnaires, in addition to biological samples, for instance to help evaluate the association between genetic variation, environmental exposures for risk of disease. Disease-specific biobanks have also been developed. One such at Barts Cancer Institute is the Pancreatic Cancer Research Fund Tissue Bank (PCRFB, for further details see <https://www.thepancreastissuebank.org>). Set up in 2016, it has collected:

- Blood samples from 2,200 consenting patients who underwent biopsies or surgery for pancreatic diseases, including pancreatic cancers (also urine, saliva and tissue samples);
- Large numbers of healthy control blood (also urine and saliva).

“Assume that Cancer Research UK is going to fund two different studies to investigate whether your research group’s biomarker panel can help stratify risk of pancreatic cancer, and enable early detection that might lead to improvements in public health outcomes. One will be a case-control study, that will report its results within 3 years and the second group they need to design a cohort study which will report its results within 13 years.”

In answering the question you will need to think about the following.

A: Case-control study with results in 3 years	B: Cohort study with results in 13 years
<p>What are your entry criteria for cases?</p> <p><i>They have to have pancreatic cancer. Think if they'd better be newly diagnosed, anyone with it (and at which stage?), or those that have died (most people with pancreatic cancer die from it).</i></p> <p>Most case-control studies match cases and controls. What would you match on and what are your entry criteria for controls?</p> <p><i>Entry criteria for controls - don't have the disease (if they did they should be a case), with no known diagnosis or past history of pancreatic cancer.</i></p> <p><i>Get a control for each case (or other reasonable ratio).</i></p> <p><i>Controls should be matched to achieve balance in important demographic characteristics (age, gender), to increase efficiency of estimation (eg. if you want to compare the utility of the biomarker in people older than 70y you need to have controls >70y, matching would ensure this)</i></p> <p><i>Remember that matching on too many things will make it difficult and overmatching can lead to cases and controls not differing enough in the exposure</i></p>	<p>What characteristics of the cohort will make the follow up easier and enable you to obtain useful results? Think about who will you recruit - what type of people, for where and how, how many</p> <ul style="list-style-type: none"> - <i>People who are easy to keep track of (so helpful to have some method of finding them if leave work) and likely to be motivated to fill forms in and come back for check-ups.</i> - <i>It needs to be large enough so that enough will get the disease in the next 13 years and/or is at high risk from other factors.</i> - <i>They need to be old enough / high risk enough so that enough will get the disease in the next 13 years and not too old because they may be quite likely to die of something else in next 15 years, so probably something like 50-79 years old.</i> - <i>They need to come from backgrounds and ethnicities representative of the target population.</i>

<p>Describe how you will recruit to the case control study (remember you can use a biobank)</p> <p><i>Retrospectively collected samples from a biobank (controls can be either healthy individual registered in the biobank, or individuals who are registered in the biobank as subjects with various diseases that are not pancreatic cancer). Or (prospective) consent new patients with pancreatic cancer from hospital (cases), and controls in the same hospitals, or at least geographical region from where the cases were recruited, or healthy volunteers, blood donors, ...</i></p>	<p>What is a simple binary risk factor or exposure?</p>
<p><i>Biomarker positive/negative at diagnosis¹.</i></p> <p>How else could you measure the risk factor (biomarker)?</p> <p><i>The biomarker might be a continuous variable, not binary. For example, gene expression could be categorised as high vs low, or the raw intensity used as a biomarker. Using a continuous variable is almost always more informative than a categorical version of the same marker.</i></p>	<p><i>Biomarker positive/negative prior to diagnosis².</i></p> <p>Briefly describe how you would determine the value of biomarker and other risk factors.</p> <p><i>Test samples on everyone at entry, and potentially also through time (additional samples and questionnaires).</i></p>

¹Ideally biomarker positivity is pre-defined, perhaps based on a cutpoint if the underlying biomarker is a continuous variable.

²as before.

<p>Give the single most accurate and useful simple binary measure of outcome for your study.</p> <p><i>Pancreatic cancer</i></p>	<p>Death from pancreatic cancer or diagnosis of pancreatic cancer - because pancreatic cancer has such a poor prognosis there is little difference. Death from cancer may be a better outcome for tracking, because even if someone is lost to follow-up we can usually still find out if they have died and from what cause using death certificates. For other cancers, death from the cancer may be a stronger outcome than disease incidence because of overdiagnosis (but not covered in this course).</p>
<p>Who will the actual results be representative of?</p> <p><i>If the case-control study is well designed, it would be representative of the study population that cases and controls are supposed to have arisen from - from where the cases and controls came.</i></p>	<p>The social and demographic make up of members of the cohort.</p>
<p>Who will these results be useful for?</p> <p><i>Because RR/HR and OR are generally robust to different populations - to everyone who might benefit of a screening program, so eventually for everyone.</i></p>	<p>Who will these results be useful for?</p>
<p>What other results could you get from your study group?</p> <p><i>Could search for other different risk factors to see if associated with pancreatic cancer.</i></p>	<p>Could ask or test about different risk factors to see if associated with pancreatic cancer and could follow up for different types of cancer or causes of death</p>

<p>Could the age of the participants affect the interpretation of your results? If so how?</p> <p>No because you have matched for age.</p>	<p>Yes, if age is correlated with the biomarker, but the biomarker does not add information beyond age (age is a confounding factor). For this reason one would like to adjust for age in the analysis.</p>
<p>What biases could occur from your study design?</p> <p>Main bias is inappropriate controls - selection bias. For instance, if controls are selected for inclusion using a criteria that is linked (correlated) with the biomarker. Confounding for other risk factors Participants with cancer were not detected at an early stage - not asymptomatic; to understand performance for early detection would ideally have a sample from cases prior to diagnosis. Will require additional studies.</p> <p>Give an example of a study of an exposure and a disease that would be unsuitable for your type of study but suitable for the other. Explain the characteristics of the exposure or disease that make it unsuitable</p> <p>A rare exposure that is associated with the disease but not the main cause of it. Would be unlikely to get anyone in the case control with the exposure.</p>	<p>Not many - those in the cohort may be a selected (self or otherwise group) but the effect of having a biomarker or not is unlikely to be different in them than in those who were not asked or declined to take part. Bias may occur due to loss to follow-up but obtaining death certificates would overcome this.</p> <p>Rare diseases - even if you have a large cohort it is unlikely that enough people can be followed up.</p> <p>Summarise the advantages in general of your type of study?</p> <p>Cheaper than cohort. Relatively quick. Good for rare diseases.</p> <p>Generally less biased than case control studies. Can estimate risks and difference in risks as well as relative risks. Can look at more than one outcome.</p>

Summarise the disadvantages in general of your type of study?	
<i>Can't do for rare exposures Suffers from confounding (other than the variables matched on) In general, selection bias Possible poor controls.</i>	<i>It is time consuming and expensive. Suffers from confounding.</i>

Table 1: Fill in the side corresponding to your group.

Further reading:

- Early detection of pancreatic cancer: <https://pubmed.ncbi.nlm.nih.gov/32135127/>
- Description of the UROPANC trial: <https://www.bartscancer.london/centre-for-experimental-cancer-medicine/clinical-trial-portfolio/uropanc/>
- A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7758047/>
- Germline BRCA2 K3326X and CHEK2 I157T mutations increase risk for sporadic pancreatic ductal adenocarcinoma: <https://onlinelibrary.wiley.com/doi/10.1002/ijc.32127>
- Noninvasive Diagnosis of Pancreatic Cancer Through Detection of Volatile Organic Compounds in Urine: <https://pubmed.ncbi.nlm.nih.gov/29129714/>
- Biomarkers in molecular medicine: cancer detection and diagnosis: <https://www.future-science.com/doi/10.2144/05384SU04>
- Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests: <https://jamanetwork.com/journals/jama/fullarticle/191668>
- Statistical Hypothesis Testing versus Machine Learning Binary Classification: Distinctions and Guidelines: <https://www.sciencedirect.com/science/article/pii/S2666389920301562>