

## Seminar 4: lung cancer

Start Assignment

- Due Sunday by 23:59
- Points 1
- Submitting a file upload
- File types pdf, r, and rmd

### General Instructions

Written solutions to all the tasks must be submitted before the **deadline: Sunday December 7 at 23:59** (as pdf files in Canvas, preferably including the most important parts of your code).

You are also expected to prepare an oral presentation of your solution for each of the tasks. The presentation should aim at taking 15 minutes, to leave room for questions and discussion. At the seminar, a member of your group (randomly chosen by the teacher) will be asked to present the solution.

We expect each of you to spend around 15h on these tasks; please plan your time and meetings keeping this in mind.


Please note that the course book (and the lectures) are not your only sources of information. There are lots of information available in the research/educational literature and on the Internet about statistics as well as R and its possibilities. You are very much encouraged to try and find methods not mentioned in the book or the lectures.

### Background

According to the World Health Organization, lung cancer is the leading cause of cancer deaths globally. Lung cancer is typically diagnosed in late stages of the disease, limiting treatment options and negatively impacting survival. Early detection of lung cancer could improve treatment outcomes. The use of healthcare data together with statistical modelling and/or machine learning have shown much promise in facilitating early detection of disease based on the medical history of the patient and biochemistry data. You are tasked with carrying out the necessary clinical evaluations to enable implementation of such a model for clinical decision-support.

## Task 1: detection of lung cancer in primary care

You are designing a clinical study to evaluate a statistical model for lung cancer screening in the primary care setting. You have identified the control as the ability of general practitioners in primary care to accurately refer suspected cases of lung cancer to specialist care for further diagnostic testing. You therefore carry out a research study using electronic health record data to investigate the relationship between physician agreement in referral practices.


The results of the study are given in the file, [data\\_t1.csv \(https://canvas.kth.se/courses/55919/files/9827434?wrap=1\)](https://canvas.kth.se/courses/55919/files/9827434?wrap=1)  ([https://canvas.kth.se/courses/55919/files/9827434/download?download\\_frd=1](https://canvas.kth.se/courses/55919/files/9827434/download?download_frd=1)). The data describes the referral decision made by two primary care physicians and if the patient was diagnosed with lung cancer or not. Both physicians assessed the same 100 patients.

- Investigate and analyse different aspects of the agreement.

## Task 2: Cyfra 21-1 as a biomarker of non-small cell lung cancer

Lung cancer can be categorised into two main forms, small cell lung cancer and non-small cell lung cancer. Cyfra 21-1 is a circulating tumour biomarker that has shown promise for detecting lung cancer, in particular non-small cell lung cancer. Your colleagues have developed a preliminary statistical model for detecting lung cancer, the model includes Cyfra 21-1 as a variable among others.

You would like to test the reliability of Cyfra 21-1 measures between two chemiluminescence kits from two different manufacturers that are currently in use in healthcare.

The data file, [data\\_t2.csv \(https://canvas.kth.se/courses/55919/files/9827435?wrap=1\)](https://canvas.kth.se/courses/55919/files/9827435?wrap=1)  ([https://canvas.kth.se/courses/55919/files/9827435/download?download\\_frd=1](https://canvas.kth.se/courses/55919/files/9827435/download?download_frd=1)), contains Cyfra 21-1 measures (in ng/mL) for the two kits.

- Explore methods for analysing the agreement and compare these.
- Which method is most suitable given the data?

## Task 3: carcinoembryonic antigen (CEA) as a biomarker of lung cancer

Carcinoembryonic antigen (CEA) has shown great potential as a biomarker for many forms of cancer, including lung cancer. CEA is measured using quantitative Elisa (enzyme-linked immunosorbent

assay).

The data file, [data\\_t3.csv \(https://canvas.kth.se/courses/55919/files/9827436?wrap=1\)](https://canvas.kth.se/courses/55919/files/9827436?wrap=1), [↓ \(https://canvas.kth.se/courses/55919/files/9827436/download?download\\_frd=1\)](https://canvas.kth.se/courses/55919/files/9827436/download?download_frd=1), details 100 measures of CEA across five different Elisa assays.

- Investigate the agreement of the CEA measures across the five assays.

## Task 4: evaluating a model for early detection of lung cancer

A classification model was developed for early detection of lung cancer based on primary care data. The model was then validated against an external retrospectively collected dataset of 1,000 matched patients with and without a lung cancer diagnosis.

- Examine the data, [data\\_t4.csv \(https://canvas.kth.se/courses/55919/files/9827437?wrap=1\)](https://canvas.kth.se/courses/55919/files/9827437?wrap=1), [↓ \(https://canvas.kth.se/courses/55919/files/9827437/download?download\\_frd=1\)](https://canvas.kth.se/courses/55919/files/9827437/download?download_frd=1).
- Analyse the performance of the model.
- What conclusions can you draw?
- Explore and discuss different ways of determining an appropriate threshold value considering severity of disease and the severe negative consequences of missing actual lung cancer cases.