**Master of Public Health, specialization: AI4PH**

**DESU: AI4PH**

**TU PHS-PRIM: Principles and methods of public health sciences**

This evaluation begins with some multiple choices questions and open questions about data analysis and bias. The second part focus on three articles that you can download on Ametice. You can answer the questions on this file and download it on Ametice. You will name it as follow:

Surname-first name-Homework-2-PHS-PRIM

**Some multiple-choice questions to start:**

1. **What is a confidence interval (choose the right answer(s))?**

* The confidence interval is an interval of values that contains 95% of the observed data
* **The 95% confidence interval is an interval of values that has a 95% chance of containing the true value of the estimated parameter**
* Neither
* Both

1. **How can a confidence interval be reduced?**

By collecting more data. The width of the CI is approximately propor-tional to the reciprocal of the square root of the sample size. So if you increase the sample size by a factor of four, you can expect the CI to be half as wide.

1. **The relationship between exposure to cigarettes and lung cancer is studied and found to be p<0.05. How can we conclude (choose the right answer(s))?**

* **There is a statistical link between cigarette exposure and lung cancer**
* Exposure to cigarettes increases the risk of lung cancer
* Having lung cancer increases the risk of being exposed to cigarettes
* Having lung cancer and being exposed to cigarettes are both linked to another factor

1. **What are alpha and beta risk? Which one is minimized in medicine? Why?**

**Type I error**

When there really is no difference (or association or correlation) between the populations, random sampling can lead to a difference (or association or correlation) large enough to be a statistically significant. This is a Type I error. It occurs when you decide to reject the null hypothesis when in fact the null hypothesis is true. It is a false positive.

**Type II error**

When there really is a difference (or association or correlation) between the populations, random sampling (and small sample size) can lead to a difference (or association or correlation) small enough to be not statistically significant. This is a Type II error. It occurs when you decide not to reject the null hypothesis when in fact the null hypothesis is false. It is a false negative.

If you set α to a very low value, you will make few Type I errors. That means that if the null hypothesis is true, there will be only a small chance that you will mistakenly call a result statistically significant. However, there is also a larger chance that you will not find a significant difference, even if the null hypothesis is false. In other words, reducing the value of α will decrease your chance of making a Type I error but increase the chance of a Type II error.

If you set α to a very large value, you will make more Type I errors. If the null hy- pothesis is true, there is a large chance that you will mistakenly conclude that the effect is statistically significant. But there is a small chance of missing a real difference. In other words, increasing the value of α will increase your chance of making a Type I error but decrease the chance of a Type II error. The only way to reduce the chances of both a Type I error and a Type II error is to collect bigger samples.

In medicine it is desirable to minimize alpha-risk as if you affirm “something” that turns out to be wrong, there could be consequences and you could be retained accountable for it whereas if you miss to identify “something”, nothing bad can happen. You “just” missed a chance for improvement, someone else will likely take it later.

It is better to miss the discovery of a new molecule compared to the release of a dangerous one on the market.

1. **Link the different disciplines to their definition:**

|  |  |  |  |
| --- | --- | --- | --- |
| Descriptive epidemiology | • | • | To study the determinants of diseases (etiologies, risk factors) and health behaviors |
| Analytical epidemiology | • | • | Measuring the extent, evolution over time and geographical distribution of health-related phenomena in human populations |
| Clinical research | • | • | To understand and evaluate the management of sick people (diagnosis, prognosis, therapy, prevention |

1. **Give an example of a possible action to minimize each of those four following difficulties that may arise when collecting survey data**

* Reliability of data
  + Declarative data are conditioned by the respondent's knowledge of the subject and the way they understand the questions:   
    correctly formalize the question in the most unambiguous way possible, rather exploit structured and objective data if available, make sure sampling is appropriated.
  + Measurement tools may give different results depending on the users:   
    simplify/change the procedure, use a friendlier tool, reduce the number of operators involved, standardize the training of the operators, evaluate their understanding of the procedure, control conditions that may influence operators (time of the day, workload,
* Non-recruitment
  + Stigma associated with certain diseases:   
    anonymity, non in-person contact.
  + Population difficult to reach:   
    geographically? Technology. If any, involve other professionals in direct contact with such population. Identify key points in their care pathway.

1. **Fill-up that table with the following information for each data sources: example of collectable data, the data scale (individual or collective scale), the name of a structure that collects this type of data, the initial purpose for their collection (research, policy making, funding…) and an advantage and a disadvantage of this type of database:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data source** | **Example** | **Scale** | **Structure** | **Initial collection purpose** | **advantage** | **disadvantage** |
| **cohorts/registers** | E4N | Individual or collective scale (i.e. the target could be *individuals* or *families*) | Inserm (research institutes) | Research | * Clarity of temporale sequence of exposure.s-event.s * Allow calculation of incidence, RR, risk difference, attributable proportion * Measure of dose-effect relationship * Facilitate study of rare exposure * Allow examination of multiple effects of a single exposure * Avoid Selection Bias at Enrollment: Cohort studies, especially prospective cohort studies, reduce the possibility that the results will be biased by selecting subjects for the comparison group who may be more or less likely to have the outcome of interest, because in a cohort study the outcome is not known at baseline when exposure status is established.   Nevertheless, selection bias can occur in retrospective cohort studies (since the outcomes have already occurred at the time of selection), and it can occur in prospective cohort studies as a result of differential loss to follow up. | * Not adapted for rare diseases * Study duration+++ if latency+++ * Cost and complexity * Lost to follow-up * Change of exposure status during follow-up * Ethical questions * Sample size * Huge database size requiring data management skills |
| **dedicated studies** | Case control studies | Individual | Clinical research units | Research | * Possible (mainly) if disease is rare * Possible and short even if long latency * Lower costs * Lack of lost of follow-up (but selection of surviving cases) * Lack of ethical problems   Possible to study several exposures | * Results will be quantified via poorly intuitive concepts, OR in particular * Estimation of incidence not possible * Memory and other bias can heavily influence data quality |
| **patient records** | Clinical narratives | Individual | Any hospital | Clinical cares | Usually readily available as it is already collected for administrative purposes | Non necessarily structured or digitalized |
| **administrative records** | PMSI | Individual | Hospitals/CNAM | Funding | * Structured and readily available. * Codes can be very reliable depending on the pathology. | * Biased in a research context as it is not collected for this purpose. * Different coding habitudes among different services * Not every code is reliable |
| **public data** | Base de données publique des médicaments | Collective | HAS (Haute authorité de santé) | Research/policy making | * Allows transparent research * Assessing a community’ need * Can lead to a flow of useful insights coming from 3rd parties | * Can be misused or misunderstood * Privacy can be violated unintentionally * Can be subject to an embargo period not allowing the publication of recent data * Database itself is transparent but what’s “behind” it may not hence data quality is not a guarantee |

1. **Which specific selection bias can artificially improve the result of a meta-analysis because of the way results are disseminated?**
2. **Biases are systematic errors that will limit the validity of a study. Which definition best fits each of the following categories:**

|  |  |  |  |
| --- | --- | --- | --- |
| Selection bias | • | • | Intervention of a third factor |
| Misclassification bias | • | • | The population observed by the survey differs from the target population in the constitution of the samples or in the follow-up of the groups |
| Confounding bias | • | • | Classification errors on one or more variables of interest |

1. **Give the name of a method used to control an identified confounding bias:**

Adjustment

1. **Give the name of a method used to control unidentified confounding bias:**

Randomization

1. **About representativeness and selection bias, choose the right answer(s):**

* **Extrapolation of the results of a study to the targeted population can’t be done when there is a selection bias**
* **Drawing the study sample from a database representative of the target population contribute to avoid selection bias**
* A way to reduce selection bias is to standardize data collection conditions
* Representativeness is needed to analyze the effect of an exposure on an outcome
* **Representativeness is needed to extrapolate the results of a study to the targeted population in real life**

1. **Give an example of selection bias in a case-control study**
2. **Give an example of confusion bias in a cohort study**
3. **Give an example of misclassification bias in a randomized trial**

.

**Articles (to download on AMetice)**

**Article 1: *The Safety and Short-Term Efficacy of Aliskiren in the Treatment of Immunoglobulin A Nephropathy – A Randomized Cross-Over Study***

1. **What is the rationale for this study?**
2. **What is the experimental design chosen for this study? Explain it.**
3. **What is the primary end point?**
4. **What are the advantages and constraints of this type of design compared to a parallel group trial?**

* Advantages:
* Constraints:

1. **The intermediate phase of 4 weeks between the two treatment periods is intended to:**
2. **Link the effect to its definition**

|  |  |  |  |
| --- | --- | --- | --- |
| Interaction between treatment effect and time period | • | • | The effect of treatment versus placebo is different in period 1 and period 2 |
| Period effect | • | • | The effect of treatment in period 2 is modified by the effect of placebo received in period 1 |
| Treatment effect | • | • | The difference in the change in proteinuria between the treatment and placebo arms was the same in Period 1 and Period 2 |
| Situation of non-interaction | • | • | The change in proteinuria on treatment was greater than the change on placebo in both periods 1 and 2 |
| Persistent effect which can be translated by an interaction | • | • | The change in proteinuria on placebo and on treatment is smaller in period 2 than in period 1 |

1. **The role of randomization in this type of design is (choose the right answer(s)):**

* To control the treatment effect
* To control the period effect
* To control for the possible effect of an interaction between the treatment and the period
* To allow a causal judgement to be made in the absence of an interaction between the treatment effect and the period

1. **A placebo is administered in the control arm. In the absence of additional information, it can be assumed that (choose the right answer(s)):**

* The study is mandatorily double-blinded
* The study is not blinded at all because of the need for immediate intervention in case of renal failure
* The study is not blinded at all, because blinding is not useful if the study is randomized
* Blinding, which ensures comparability between groups, is not useful if the study is randomized
* The study is certainly blinded, but we cannot say if it is single or double-blinded

1. **Among the following proposals, give the criterion that can be used to judge a causal relationship between a risk factor and a disease (choose the right answer(s)):**

* Temporality: the effect occurs after the cause
* Reproducibility
* Dose-response relationship
* The existence of a known mechanism that can explain the relationship
* No suitable proposal

**Article 2: *Deep learning for prediction of colorectal cancer outcome: a discovery and validation study***

1. **What was the aim of that study?**
2. **What kind of methodology could be used in a future study to increase the level of proof of the result? Why is that methodology increasing the level of proof?**
3. **The authors found that the DOMore-v1-CRC classifier has a sensitivity of 52%, a specificity of 78%, a positive predictive value of 19% and a negative predictive value of 94% when comparing 3-year CSS to DOMore-v1-CRC good prognosis vs uncertain and poor prognosis. Define those parameters and rephrase those results.**

* Sensitivity:
* Specificity:
* Positive predictive value:
* Negative predictive value:
* The DOMore-v1-CRC classifier has a sensitivity of 52% when comparing 3-year CSS to DOMore-v1-CRC good prognosis vs uncertain and poor prognosis:
* The DOMore-v1-CRC classifier has a specificity of 78% when comparing 3-year CSS to DOMore-v1-CRC good prognosis vs uncertain and poor prognosis:
* The DOMore-v1-CRC classifier has a positive predictive value of 19% when comparing 3-year CSS to DOMore-v1-CRC good prognosis vs uncertain and poor prognosis:
* The DOMore-v1-CRC classifier has a negative predictive value of 94% when comparing 3-year CSS to DOMore-v1-CRC good prognosis vs uncertain and poor prognosis:

1. **In general, if you develop a new test, what parameter will you favour if the disease is very severe but easily curable with a non-invasive treatment (sensitivity or specificity)? Why?**
2. **In general, if you develop a new test, what parameter will you look first if the disease is severe with a difficult treatment and a high risk of side effects (sensitivity or specificity)? Why?**

**Article 3:** ***Effectiveness of quadrivalent human papillomavirus vaccine for the prevention of cervical abnormalities: case-control study nested within a population-based screening programme in Australia***

1. **The introduction refers to “Phase III studies”. What those terms refers to? What are the other phases for drug development (purpose, population involve)?**
2. **Why the authors made the choice to focus on women who presented for their first smear test?**
3. **Some abnormalities may have gone unnoticed and undiagnosed. What kind of bias can this cause? Justify**
4. **A high level of education in the family is associated with a higher frequency of vaccination and less risky sexual behavior. What kind of bias is this?**
5. **Which ethical procedures have been implemented for this study?**
6. **What procedure has been implemented to avoid that the researchers have access to the name of the women involve in the study during the linkage of the databases?**