TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Aplicació del cicle “Problem - Plan - Data - Analysis - Conclusion” (PPDAC) en ciències de l’esport

Autor: Sara Montañés González

Director: Martí Casals Toquero i Jordi Cortés Martínez

Departament: Departament d’Estadística i Investigació Operativa

Convocatòria: Juny 2022

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Agraïments

A en Martí Casals i en Jordi Cortés, tutors d'aquest treball de final de grau, per el seu constant suport, per la confiança, per tot el coneixement compartit i fer-me gaudir tant de la realització d'aquest Treball de Final de Grau.

A les meves companyes d'universitat pel suport rebut durant aquests mesos, pel seu interès i per fer el procés menys feixuc.

A la meva família i amics pel seu inqüestionable recolzament i acompanyament.

Resum i paraules clau

L'estadística en les ciències de l'esport ha despertat un gran interès als darrers anys. Per fer front a qüestions en aquest àmbit es demanen cada vegada més científics de dades. Aquests perfils passen sovint molt de temps analitzant les dades sense tenir habilitats de pensament estadístic i computacional que poden ser rellevants en la fase de disseny d'un estudi. La metodologia PPDAC (Problem – Plan – Data – Analysis – Conclusion) creada al 1999 és una estructura cíclica que ajuda a resoldre problemes i prendre decisions a nivell quantitatiu en qualsevol àmbit de la ciència. El present treball presenta un nou instrument dividit en les cinc parts del PPDAC amb 60 preguntes construïdes a partir d'articles científics d'educació estadística. Per comprovar les possibilitats d'aquest instrument s'exposen dos exemples amb dades reals del camp de l'esport que puguin servir de model a investigadors per conèixer quines preguntes poden ser d'utilitat en el disseny, anàlisi i comunicació d'un estudi científic.

Paraules clau: PPDAC, alfabetització de dades, alfabetització estadística, raonament estadístic, pensament estadístic, cicle investigatiu

---

Aplicación del ciclo Problem – Plan – Data – Analysis – Conclusions (PPDAC) en ciencias del deporte

La estadística en las ciencias del deporte ha despertado un gran interés en los últimos años. Para dar respuesta a cuestiones en este ámbito se buscan cada vez más profesionales o científicos de datos. Estos perfiles dedican mucho tiempo al análisis de datos sin tener habilidades de pensamiento estadístico y computacional que pueden ser esenciales en la fase de diseño de un estudio. La metodología PPDAC (Problem – Plan – Data – Analysis – Conclusion) nacida el año 1999 en Nueva Zelanda es una estructura cíclica que ayuda a resolver problemas y tomar decisiones a nivel cuantitativo en cualquier ámbito científico. En este trabajo se ha creado un instrumento dividido en cinco partes del PPDAC con 60 preguntas construidas a partir de trabajos de investigación de educación estadística. Para comprobar las posibilidades de este instrumento se muestran dos ejemplos con datos reales en el campo del deporte que puedan servir de modelo a investigadores y científicos de datos para conocer qué preguntas pueden ser de utilidad antes de la toma de decisiones en sus análisis.

Palabras clave: PPDAC, alfabetización de datos, alfabetización estadística, raonamiento estadístico, pensamiento estadístico, ciclo investigativo

Aplication of the cycle Problem – Plan – Data – Analysis – Conclusions (PPDAC) in sports sciences

Statistics in sports sciences has incited a great interest in recent years. In order to deal with questions related to this field, more and more professionals or data scientist are needed. These professional profiles often spend a lot of time analysing data without having statistical and computational thinking skills. The PPDAD (Problem – Plan – Data – Analysis – Conclusion) methodology, originated in 1999 in New Zealand, is a cyclical structure that helps to solve problems and make decisions on a quantitative level in any scientific field. In this thesis, an instrument has been created divided into the five parts of the PPDAC, from a total of 60 questions created based on statistical research papers. In order to check the possibilities of this instrument, two examples are presented with real data from the field of sports that can be used as a model for researchers and data scientists to realize what questions may be useful before the decision-making of their analysis.

Key words: PPDAC**,** data literacy, statistical literacy, statistical reasoning, statistical thinking, investigative cycle

---

Classificació AMS

101C 08 Critical thinking
101C 44 Inquiry-based learning-research on inquiry methods
101C 48 Inquiry-based learning-classroom implementation
101C 52 Inquiry-based learning-resources
101B 12 Problem solving
101B 40 Quantitative literacy
104A 82 Mathematics and sports

Índex

Índex de taules

Índex de figures

1. Introducció

El reconeixement de l'estadística ha coincidit probablement amb l'ús del terme *data scientist* o científic de dades, la professió més “sexy” del segle XXI segons *Harvard Business Review* (Hayes Davenport et al., 2012). Actualment, s'estan començant a utilitzar altres termes relacionats amb l'àmbit com són *data literacy*, *statistical literacy*, *statistical reasoning* o *statistical thinking* (Jeffrey O. et al., 2003; Sabbag et al., 2018; Schield, 2017; Utts, 2014; Wild et al., 2018). De fet, l'objectiu genèric de l'estadística com a eina social és convertir les dades en coneixement del món real.

Una de les línies de recerca en estadística és l'educació estadística. L'any 2005, l'Associació Nord-americana d'Estadística (ASA) va aprovar unes guies (GAISE) per a l'avaluació i l'ensenyança a l'educació estadística a la universitat (Aliaga et al., 2005). Aquest informe ha tingut un impacte profund en l'ensenyament de la introducció de l'estadística i implica un canvi de rumb en la forma de plantejar l'estadística en el futur basant-se en sis recomanacions. Una d'elles es centra a ensenyar *Statistical thinking*, terme que es pot entendre com l'encarnació estadística del sentit comú, ja que, la seva presència molts cops no és visible, però la seva absència clarament ho és (Wild et al., 1999). Per poder desenvolupar la capacitat de pensament estadístic, cal ensenyar estadística com un procés de recerca, de resolució de problemes i de presa de decisions que es pot dur a terme gràcies al cicle PPDAC (Problem – Plan – Data – Analysis – Conclusion) que dotarà als estudiants o a la societat d'experiència amb el pensament multivariable. El PPDAC, terme originari de Nova Zelanda i tractat per primer cop per l'estadístic C.J. Wild i el matemàtic M. Pfannkuck, va ser dissenyat per marcar els passos a seguir a l'hora de resoldre un problema fent servir evidència científica a nivell quantitatiu (Wild et al., 1999). En l'article es discuteix el procés de resolució de problemes estadístics, creant una estructura formada per quatre dimensions. La primera d'elles, anomenada *investigative cycle*, va acabar esdevenint al PPDAC i una part important d'aquesta consisteix a comprendre que resoldre un problema implica solucionar una mancança de coneixement. La segona dimensió, anomenada com *types of thinking* enumera diferents tipus de pensament, dividint-ho en pensaments generals i pensaments fonamentals per a l'*statistical thinking*. La dimensió 3, anomenada *investigative cycle*, tracta sobre el procés de pensament genèric en la resolució de problemes, dividint-ho en cinc components: *Generate, Seek, Interpret, Criticise i Judge*. Finalment, la quarta dimensió, anomenada *Dispositions*, discuteix diverses qualitats que un científic hauria de tenir i que afecten l'*statistical thinking.*

Un dels camps d'aplicació de l'estadística amb major creixement els darrers anys ha estat l'esport gràcies a múltiples factors: 1) les constants innovacions tecnològiques que permeten la recollida de grans quantitats de dades sobre posicionament de jugadors; 2) l'aparició de pel·lícules com *Moneyball* que mostren l'ús indispensable de l'estadística en el beisbol; 3) l'aparició de conferències obertes d'anàlisi de l'esport; o 4) recents treballs de científics de dades de l'esport en revistes d'estadística o editorials reconegudes per l'*American Statistical Association* (ASA) (Statistical thinking in sports, 2007; Cervone et al., 2016; M. J. Lopez et al., 2018, 2015; Macdonald, 2020). L'estadística aplicada a ciències de l'esport és adequada per ser emprada en l'àmbit esportiu a causa de la disponibilitat de dades, el coneixement preexistent dels esports entre els estudiants i la seva societat i la facilitat amb què es poden utilitzar per promoure l'alfabetització de dades. Les nombroses fonts públiques de dades disponibles fan que la resolució de problemes del món real sigui una opció que no està disponible a molts altres camps. Aquestes característiques fan que els exemples relacionats amb l'anàlisi de l'esport siguin una opció excel·lent per a les introduccions no tècniques al raonament i al pensament estadístic (Starkings et al., 2008). Tot i que darrerament hi ha diferents treballs fent èmfasi de la importància de l'estadística, la seva correcta utilització i els seus professionals en aquest àmbit (Bullock et al., 2022; Lopez et al., 2018; K. Sainani et al., 2022; K. L. Sainani et al., 2021), fins on nosaltres coneixem, només hi ha un llibre que es centra en una visió general dels mètodes estadístics en l'esport i tractament en profunditat dels problemes i reptes crítics als quals s'enfronta la investigació estadística en aquest camp (Albert et al., 2017). Concretament, la resolució de problemes reals a partir del PPDAC en aquest àmbit pot ser crucial tant si són a curt, mitjà o llarg termini.

Actualment, disposem de moltes dades procedents de diferents esports i aplicades per propòsits diversos, però poques vegades s'intenta utilitzar el pensament estadístic per a la resolució de problemes en el camp de l'esport. És per això que l'objectiu del present treball és ensenyar els avantatges d'utilitzar el cicle PPDAC en l'àmbit de l'esport mitjançant exemples pràctics i reals. També es proposarà un nou instrument basat en la metodologia PPDAC per avaluar l'ús del pensament crític en les diferents fases d'un estudi (disseny, anàlisi i comunicació de resultats). S'aplicarà aquesta metodologia i aquest instrument a dos casos d'estudi del món de l'esport.

2. Mètode

2.1. El PPDAC

La societat sempre ha volgut comptar i mesurar tot allò que l'envolta, intentant així entendre el món en què habita. L'estadística moderna però no va esdevenir una disciplina fins voltants l'any 1650 i va ser més endavant, al segle XX, quan l'estadística va començar a tenir una base més matemàtica i per a molts estadístics va acabar esdevenint una "motxilla d'eines estadístiques" (Spiegelhalter, 2019) que es basava en receptes per poder resoldre problemes científics o quantitatius.

Aquesta pobra visió de l'estadística està actualment enfrontant-se a altres reptes, els quals consisteixen en no només dur a terme anàlisis estadístiques, sinó també en entendre i ser crítics amb les conclusions extretes de qualsevol anàlisi. Aquesta idea podria ser englobada en el terme *data literacy* (alfabetització de dades). Aquest, exigeix que l'ensenyament de l'estadística estigui més enfocat a la resolució de problemes, on l'aplicació de l'esmentada "motxilla d'eines estadístiques" sigui només una part d'aquest procés i així el PPDAC (Problem – Plan – Data – Analysis – Conclusions) sigui una estructura cíclica pensada amb aquesta finalitat que va més enllà del mètode científic.

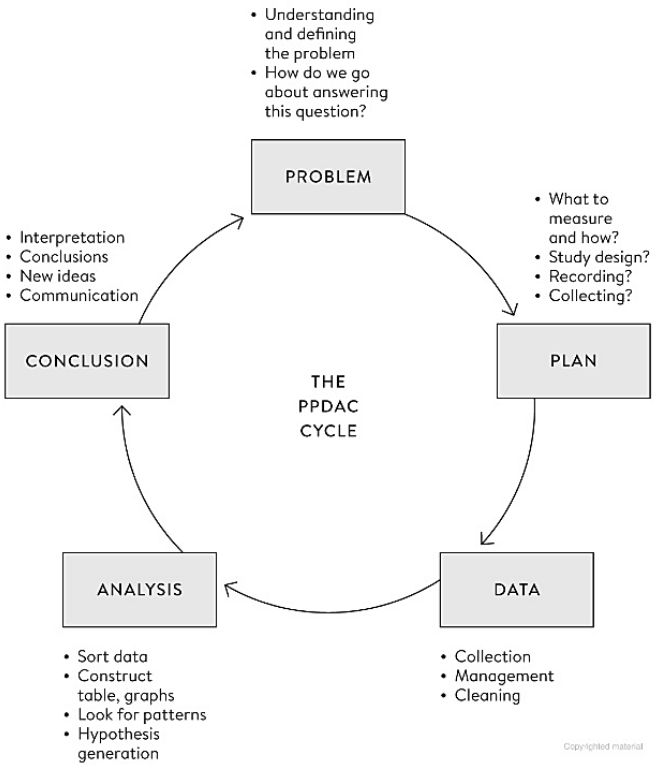

Figura 2.1. El cicle PPDAC actualitzat per David Spielgelhalter

Actualment, molts dels investigadors/es o usuaris/es que volen enfrontar-se a un problema gasten molt de temps en el *Data* i *Analysis*, sense prendre la suficient atenció a la base del problema que es basa justament en el *Problem* i *Plan*.

El primer pas de la metodologia del cicle PPDAC rep el nom de *Problem*, i consisteix en conèixer i definir el problema o pregunta al qual es vol donar resposta. Un cop definit el problema es planifica (*Plan*) què es mesurarà i com per poder donar resposta a la pregunta plantejada en l'apartat anterior. En aquesta secció també es pensa el disseny de l'estudi i com seran les dades que s'han de recollir per poder dur a terme l'anàlisi. Tot seguit, en el tercer pas del cicle (*Data*) es recullen les dades, es processen i se'n comprova la seva qualitat. Quan ja es tenen les dades es procedeix a analitzar-les en l'apartat *Analysis*, aplicant allò que s'ha plantejat en el segon pas. Per acabar, l'últim pas d'aquesta metodologia rep el nom de *Conclusions*, i consisteix en comunicar de manera clara i entenedora els resultats que s'hagin obtingut durant l'aplicació del cicle, tenint en compte les possibles limitacions que es tenen i sent crítics amb allò que s'ha obtingut. En general les conclusions fan sorgir noves preguntes, fent que el cicle pugui començar de nou i així millorar la seva resposta.

2.2. Instrument *checklist*

S'ha construït una llista de comprovació (*checklist*) dividida en les cinc seccions del PPDAC on es proposen preguntes que un/a investigador/a hauria de respondre a l'hora de resoldre un problema, independentment de la temàtica d'aquest. Totes les preguntes que s'han inclòs es basen en articles científics relacionats amb el pensament estadístic, l'educació estadística i el *data literacy*. En la taula cada pregunta disposa d'una breu explicació més en detall per comprendre l'objectiu de la qüestió. Les etiquetes de les respostes a cadascuna de les subpreguntes són: Sí, No, No està clar i No procedeix. L'objectiu de l'investigador es obtenir el major nombre de "Sí" dintre de les preguntes que són procedents.

En general existeixen dos tipus de taules de comprovació (*Checklist*): aquelles que donen consells sobre com procedir en un tipus d'estudi, com per exemple la *Consort checklist* (Moher et al., 2012) o d'altres que atorguen una puntuació per avaluar algun aspecte, com per exemple la *GRADE* (Guyatt et al., 2008), que avalua el nivell d'evidència científica d'un estudi. En aquest treball farem servir una taula de

comprovació amb el mateix objectiu que el primer cas, la *Consort checklist*, ja que el segon tipus requereix una validació més profunda.

L'instrument creat es pot trobar en l'Annex del present document i està dividit en les següent seccions:

1. Problem
    a. Comprensió i definició del problema
2. Plan
    a. Què mesurar i com?
    b. Disseny de l'estudi
    c. Recollida de dades ("Collecting")
    d. Passos-cicle i pensament computacional
    e. Enregistrament de les dades ("Recording")
3. Data
    a. Supervisar/Auditar les dades
    b. Maneig de les dades
    c. Netejar les dades ("Cleaning")
4. Analysis
    a. Classificació de les dades ("Sort of data")
    b. Tècniques estadístiques utilitzades
    c. Construcció de taules i gràfics
5. Conclusions
    a. Interpretació
    b. Conclusions
    c. Comunicació
    d. Impacte

Cal mencionar que aquest instrument construït no estaria només pensat per a estadístics, sinó també per a usuaris o investigadors/es que volen resoldre problemes amb un rigor més científic i sense oblidar el pensament estadístic.

3. Aplicació

En aquesta secció s'estudiaran dos casos, un de futbol i un de bàsquet, seguint la metodologia del PPDAC i fent servir la llista de comprovació construïda. En ambdós casos el problema a resoldre s'establirà a partir de les dades, fet que no es correspon amb una situació real on la recollida de les dades es deriva de l'objectiu de l'estudi. No obstant, els exemples presentats són útils per il·lustrar l'ús de l'instrument.

El codi amb l'anàlisi dels dos estudis de cas es pot trobar en un repositori Github (https://github.com/saramontanes/PPDAC).

3.1. Estudi de cas 1

3.1.1. Resum executiu

Les dades d'aquest cas disposen d'informació referent als partits de futbol de la lliga italiana (*Serie A*) durant la temporada 2015/16. Per a cada jornada es disposa de diversa informació tant de l'equip local com el visitant: nombre de gols en aquella jornada, mitjana de gols en les últimes 5 jornades, posició de l'equip a la lliga en la temporada anterior, etc.
En la taula 3.1. es mostra un resum bàsic de la temàtica de les dades i dels anàlisis que es duran a terme.

Taula 3.1. Resum de les dades de l'estudi del cas 1

Disseny de l'estudi	Temàtica	Mètode estadístic
Cohorts retrospectiu	Anàlisi de futbol	Models lineals generalitzats (GLM) *Bradley-Terry*

Origen de les dades

La base de dades ha estat obtinguda a través del paquet `engsoccerdata` d'R CRAN (*Comprehensive R Archive Network*). El paquet, creat per James Curley l'any 2016 i titulat *English and European Soccer Results 1871-2016* permet accedir a diversos jocs de dades de resultats de futbol de lligues angleses i europees, entre ells el que es farà servir en l'apartat que ens ocupa (P. Curley, 2016).

Antecedents

Una aplicació de l'estadística i d'anàlisi de dades molt comú en el món de l'esport és la predicció dels resultats d'un partit o competició. Aquests resultats són de gran interès en la indústria de l'esport (per exemple en clubs, federacions o organitzacions, staff...) però també a nivell acadèmic, on es pot aprofitar per exemple aquesta informació per fer noves estratègies d'estils de joc o a l'hora de fer noves contractacions de jugadors. Les cases d'apostes també aprofiten aquests resultats i altres tecnologies aplicades a l'esport per desenvolupar la seva activitat.

Objectiu

En aquest estudi es vol predir quin equip guanyà un partit concret, sabent quin és l'equip local i l'equip visitant.

3.1.2. Resolució del cas seguint el PPDAC

Problem

Per als equips tècnics i sobretot per alguns professionals de l'esport (exemple: analistes, entrenadors, mànagers,...) dels clubs de futbol és de gran interès poder saber si el seu equip guanyarà un partit. Tot i no ser un problema nou en aquest àmbit sí que és un problema recurrent en aquest camp i aprofitant les evidències, metodologies i pensament crític actual es vol predir quin equip guanyarà un partit concret, sabent quin és l'equip local i l'equip visitant.

Plan

Per fer-ho es recolliran dades sobre partits de futbol d'una temporada i lliga concreta. Per a cada partit es tindrà el nom dels equips que el disputen, així com els gols realitzats per cada un d'ells.

Un cop obtingudes les dades es crearan dos **models lineals generalitzats**, un per predir el nombre de gols de l'equip local i un altre per predir el nombre de gols de l'equip visitant. Es començarà per aquesta opció més simple i s'avaluarà si aquesta ja proporciona resultats satisfactoris abans de provar altres opcions més sofisticades com, per exemple, ajustar un model Poisson Bivariant (AlMuhayfith et al., 2016).

Els models lineals generalitzats (MLGz) tenen les següents components:

- Component **determinista**, esperança de $Y = (Y_i)$ donada per:
  - **Predictor lineal**: $\eta_i = (X_{i,1}, \dots, X_{i,K})\beta$, en global $\eta = X\beta$.
  - **Funció d'enllaç** (*link*): és la funció (bijectiva) que relaciona el valor esperat μ_i amb el predictor lineal $\eta_i, g(\mu_i) = \eta_i$, és a dir, $g(\mu) = \eta = X\beta$ En conseqüència tindrem que $\mu_i = g^{-1}(\eta_i)$ i globalment quedarà $\mu = g^{-1}(\eta) = g^{-1}(X\beta)$.

- **Component aleatòria**, per cada Y_i:
  - La funció de densitat és $f_{Y_i}(y;\ \theta_i, \phi) = exp\left(\frac{\theta_i y - b(\theta_i)}{a(\phi)} + c(y, \phi)\right)$
    - Les Y_i han de ser independents
    - θ_i és el paràmetre canònic i és funció de $\mu_i, \theta_i = \theta(\mu_i)$
    - $\Phi = a(\phi)$ és el paràmetre de dispersió i $\sqrt{\Phi}$ és el paràmetre d'escala. Per a totes les Y_i tenen el mateix valor.

La diferència dels models lineals generalitzats respecte dels models lineals generals és que admeten altres distribucions diferents a la Normal.

El model lineal generalitzat que s'aplicarà serà un que consideri una variable resposta de comptatges, ja que la variable "nombre de gols" compleix les següents propietats que també compleixen els models de recomptes:

- El nombre de gols són no negatius
- El nombre de gols no estan acotats superiorment
- En el nombre de gols el valor zero té un percentatge no menyspreable

Existeixen diversos models de recomptes en funció de la distribució de les observacions: model Poisson, Quasipoisson, Binomial negatiu, Zero-inflat... Per començar es plantejarà un model de Poisson i, a continuació, a través de l'anàlisi de la dispersió es decidirà si aquest és l'adequat o si cal canviar el model. Una altra opció seria implementar un model Poisson Bivariant, però en aquest cas no es farà així.

Altres models que també es podrien implementar i que no són de recomptes són els models Hurdle, els quals estan formats per dues parts: la primera d'elles és la probabilitat d'obtenir el valor 0, i la segona part és la probabilitat d'obtenir valors diferents de 0. És un tipus de model molt semblant als models zero-inflats.

Data

Per crear aquest model de predicció s'importen les dades de la lliga Italiana *Serie A* des de l'any 1929 fins a l'any 2015. Les dades es troben en el paquet `engsoccerdata` d'R CRAN i estan formades per 25.404 observacions, corresponents a 25.404 jornades o partits de futbol. La base de dades està depurada i no hi ha valors faltants ni valors atípics. Consta de 8 variables, que es mostren a la següent Taula 3.2:

Taula 3.2. Descripció de la base de dades de l'estudi de cas 1

	Variable	Etiquetes/rang	Descripció
1	Date	06/10/1929 – 15/05/2016	Data en què es va disputar cada partit
2	Season	1929 – 2015	Temporada (any d'inici) en què es va disputar cada partit
3	home		Nom de l'equip local
4	visitor		Nom de l'equip visitant
5	FT	0-0, 0-1, ...	Resultat (en gols) de cada partit
6	hgoal	0 – 10	Nombre de gols marcats per l'equip local
7	vgoal	0 – 8	Nombre de gols marcats per l'equip visitant
8	tier	1	Categoria de la lliga (primera divisió)

Abans de procedir en la realització de l'anàlisi es modifica la base de dades, filtrant la variable *Season* per realitzar l'anàlisi només amb les dades de la temporada 2015. D'aquesta manera el nombre d'observacions es redueix a 380. Es decideix escollir aquesta temporada perquè és de la última que es disposa de dades.

Tot seguint es creen les variables que es mostren a la Taula 3.3:

Taula 3.3. Descripció de les variables afegides a la base de dades de l'estudi de cas 1

	Variable	Etiquetes/rang	Descripció
1	journey	1 – 38	Jornada de la temporada en que es va disputar el partit
2	hpos2014	1 – 17	Posició final en la classificació la temporada anterior (2014/15) per a l'equip local

3	hgoal2014	28 – 72	Nombre de gols realitzats en la temporada anterior (2014/15) per a l'equip local
4	vpos2014	1 – 17	Posició final en la classificació la temporada anterior (2014/15) per a l'equip visitant
5	vgoal2014	28 – 72	Nombre de gols realitzats en la temporada anterior (2014/15) per a l'equip visitant
6	hpoints	0 – 3	Nombre de punts aconseguits per l'equip local en aquella jornada
7	vpoints	0 – 3	Nombre de punts aconseguits per l'equip visitant en aquella jornada
8	win	-1, 0, 1	Pren el valor 1 si l'equip local guanya, valor 0 si el resultat del partit és un empat i -1 si el resultat de l'equip és favorable per a l'equip visitant
9	hgoal_prev5	0 – 3.80	Mitjana de gols marcats per l'equip local en les últimes 5 jornades
10	hpoints_prev5	0 – 3	Mitjana de punts obtinguts per l'equip local en les últimes 5 jornades
11	vgoal_prev5	0 – 5	Mitjana de gols marcats per l'equip visitant en les últimes 5 jornades
12	vpoints_prev5	0 – 3	Mitjana de punts obtinguts per l'equip visitant en les últimes 5 jornades

Per a la creació de les variables *hpos2014*, *hgoal2014*, *vpos2014* i *vgoal2014* s'han buscat els resultats de la temporada anterior (2014/15) per a la mateixa categoria de la lliga italiana. Així doncs, els equips que la temporada anterior no competien en aquesta mateixa categoria no tenen valors en aquestes variables.

Les variables *hgoal_prev5*, *hpoints_prev5*, *vgoal_prev5* i *vpoints_prev5* prenen el valor 0 per a la primera jornada, doncs no hi ha dades de jornades anteriors per calcular la mitjana. Per a les jornades 2, 3 i 4 la mitjana es calcula tenint en compte els 2, 3 i 4 primers partits disputats.

Així doncs, la base de dades resultant consta de 20 variables: les 8 originals i les 12 variables creades a partir de les originals.

Analysis

Tal com s'ha indicat en l'apartat previ es plantejarà inicialment un model de Poisson. Per fer-ho, es dividirà la base de dades en dues parts: la mostra d'entrenament (*traint*) o de calibració, que correspondrà a dos terços de les dades (66.66%) i que es farà servir per ajustar el model, i la mostra de validació (*test*), amb el percentatge restant, sobre la qual s'avaluarà la capacitat predictiva del model ajustat amb la mostra d'entrenament. Aquesta divisió no serà arbitrària, sinó que es farà de manera no aleatòria (cronològica, en aquest cas) seguint la guia TRIPOD (Collins et al., 2015a).

Es crea el model per predir els gols de l'equip local fent servir la funció `glm` d'R, indicant `family = poisson`. Es vol predir la variable `hgoal` (variable resposta) a partir de les variables `home`, `visitor`, `hpoints_prev5`, `vpoints_prev5`, `hgoal_prev5` i `vgoal_prev5`. A continuació es fan les prediccions amb la funció `predict` i s'obtenen valors d'entre 0.39 i 3.96 gols (Taula 3.4).

Taula 3.4. Resum numèric de la predicció dels gols per a l'equip local

Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.39	1.01	1.29	1.49	1.91	3.96

A continuació es crea el model per predir els gols de l'equip visitant, fent servir les mateixes predictores però en aquest cas intentant predir la variable `vgoal`. S'obtenen valors d'entre 0.28 i 2.77 (Taula 3.5).

Taula 3.5. Resum numèric de la predicció dels gols per a l'equip visitant

Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.28	0.80	1.06	1.14	1.47	2.77

Un cop realitzada la predicció es mira si els dos models creats fan bones prediccions. Per fer-ho, primer de tot es fa un gràfic col·locant en l'eix de les abscisses el nombre real de gols i en les ordenades la predicció del nombre de gols.

En el cas del model per a l'equip local, s'obté el següent gràfic de valors reals vs. valors predits. S'observen molts valors llunyans a la bisectriu, indicant que el model no prediu bé.

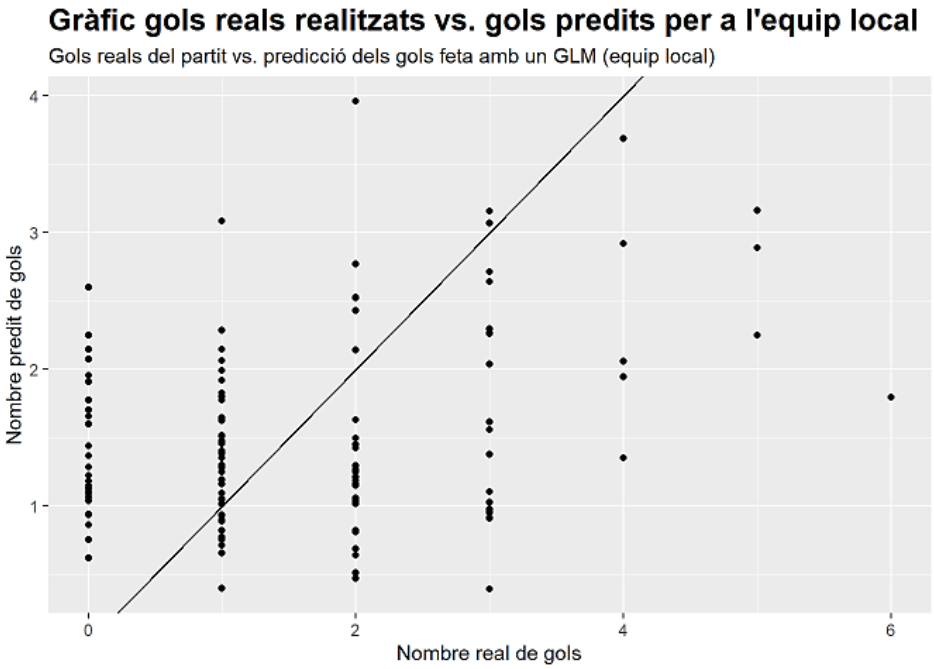

Figura 3.1. Gràfic dels valors reals vs. valors predits per a l'equip local

Per comprovar-ho es calculen els valors de la desviació típica i de l'arrel de l'error quadràtic mitjà (RMSE). La Taula 3.6 mostra que els valors obtinguts són força similars, inclús el RMSE és superior. Això implica que el model no redueix la incertesa en el nombre de gols i, per tant no aporta gaire informació a l'hora de fer les prediccions.

Taula 3.6. Valors de la desviació típica i el RMSE per al model que prediu els gols de l'equip local

Desviació típica	RMSE
1.22	1.25

Es fa el mateix anàlisi per al model que prediu els gols per a l'equip visitant. A la figura 3.2 es veuen molts punts (valors reals) que es troben molt allunyats de la línia (valors predits), de manera que aquest model tampoc fa bones prediccions.

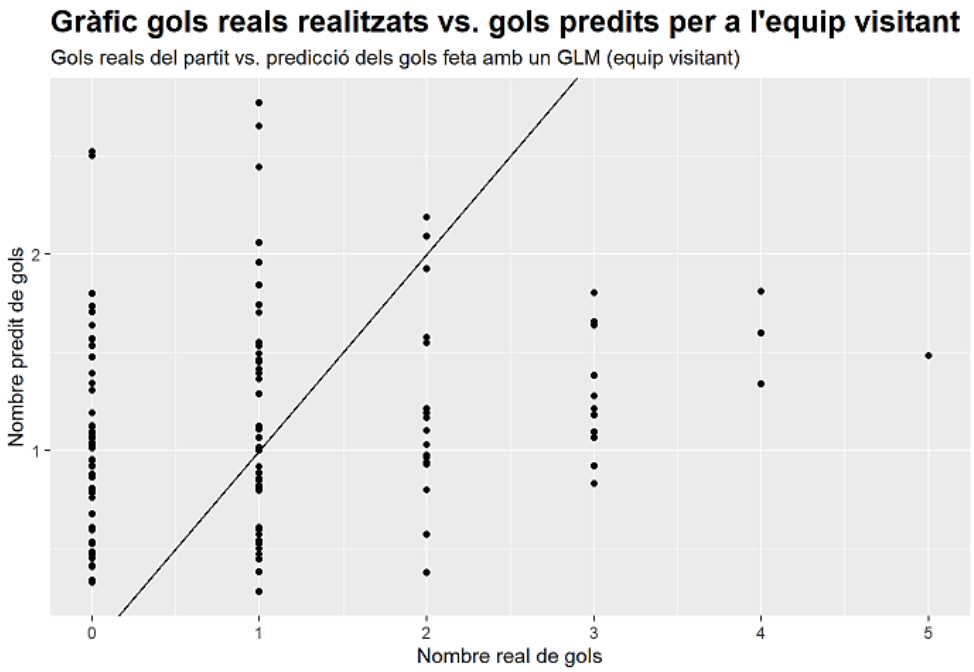

Figura 3.2. Gràfic dels valors reals vs. valors predits per a l'equip *visitor*

El càlcul de la desviació típica i del RMSE confirmen que el model fa prediccions dolentes, ja que a la Taula 3.7 veiem que prenen valors molt similars.

Taula 3.7. Valors de la desviació típica i el RMSE per al model que prediu els gols de l'equip visitant

Desviació típica	RMSE
1.08	1.12

Conclusions

A causa dels resultats obtinguts es torna al ***Plan*** per intentar donar resposta al ***Problem*** d'una altra manera. Les prediccions aconseguides amb els dos models lineals generalitzats no són bones i, a més a més, els models plantejats consideren que són independents els gols de l'equip local i l'equip visitant, i no contemplen un factor bàsic en els partits (sigui de futbol o de qualsevol altre esport col·lectiu), que és el fet de jugar en el propi camp on el del rival.

Aprofitant que el PPDAC és cíclic es tornarà a realitzar l'anàlisi fent un plantejament diferent després d'avaluar quina de les seccions del PPDAC pot tenir possibilitats de millora, i tenint en compte totes aquestes mancances presents en els models lineals generalitzats plantejats.

Plan (2) i Data (2)

En aquest cas s'implementarà un model de *Bradley-Terry* (BTM), metodologia molt utilitzada en anàlisi dels esports quan es tracten casos d'un participant versus un altre. Alguns aspectes d'aquest model són:

- S'assumeix que hi ha dos equips que es representen amb els subíndexs i, j

- Les odds que l'equip i guanyi a l'equip j es poden calcular com

$$\text{Odds(i guanya j)} = \frac{\text{P(i guanya j)}}{\text{P(j guanya i)}} = \frac{\alpha_i/(\alpha_i + \alpha_j)}{\alpha_j/(\alpha_i + \alpha_j)} = \frac{\alpha_i}{\alpha_j}$$

  on α_i i α_j s'anomenen *abilities* i es poden estimar a partir dels resultats de partits anteriors. Aquestes *abilities* es calculen fent $\alpha_i = exp\ (\lambda_i)$, on λ_i és entès com la força (o *strength*) de l'equip i. Dins d'aquest paràmetre λ_i s'hi poden incloure diversos factors que podrien afectar el resultat del partit. En el cas que ens ocupa, l'expressió de λ_i serà:

$$\lambda_i = \beta h_i$$

  On $h_i = 1$ si l'equip *i* juga a casa i $h_i = 0$ en cas contrari. L'únic paràmetre a estimar és β, el qual pot ser entès com la diferència de *strengths* quan l'equip juga en el seu propi camp (Tsokos et al., 2019).

- El model es pot expressar com un model lineal logit (Firth, 2005)

$$\text{logit}[\text{P(i guanya j)}] = \lambda_i - \lambda_j$$

- La probabilitat que l'equip i guanyi a l'equip j s'expressa com

$$\text{P(i guanya j)} = \frac{\alpha_i}{\alpha_i + \alpha_j}$$

Així doncs, per poder donar resposta al problema plantejat només és necessari calcular les *abilities* i, j (M. Lopez et al., n.d.; Skidmore College, n.d.-b).

S'aprofitaran les dades recollides i processades anteriorment, només canviant la variable `win`, fent que aquesta prengui el valor 1 si guanya l'equip visitant, 0 si guanya l'equip amfitrió i 0.5 si el resultat del partit és un empat.

És important destacar que en el model original de *Bradley-Terry* (B-T) implementat a R no hi ha la possibilitat d'empatar (Ley et al., 2019). Tot i així hi ha l'opció de codificar la variable `win` tal com s'ha especificat anteriorment, és a dir, donant el valor 0.5 a la opció d'empatar (Whelan et al., 2021).

El model original a R tampoc contempla l'avantatge que té l'equip local pel fet de jugar en el seu propi camp, no només degut al suport de l'afició, sinó també pel fet que l'equip local té més experiència en aquell camp. Per tenir aquest aspecte en compte s'afegeix l'argument `at.home` al model, el qual pren el valor 1 si l'equip juga en el seu propi camp i el valor 0 en cas contrari (Skidmore College, n.d.-a). Aquest argument fa que l'expressió principal per al càlcul de les probabilitats es vegi afectada de la següent manera:

$$\mathrm{P}(\text{i guanya j} \mid \text{i juga a casa}) = \frac{\theta\lambda_i}{\theta\lambda_i + \lambda_j}$$

$$\mathrm{P}(\text{i guanya j} \mid \text{j juga a casa}) = \frac{\lambda_i}{\lambda_i + \theta\lambda_j}$$

On el paràmetre $\theta > 1$ representa l'avantatge de jugar a casa. Com més gran sigui aquest valor, més probable és que l'equip local guanyi (Kingsman, 2016).

Analysis (2)

Per crear el model de *Bradley-Terry* es carrega el paquet `BradleyTerry2` a R CRAN (Turner et al., 2012). De nou, se separa el joc de dades en dos: la part *train* i la part *test*.

S'estima el model amb la funció `BTm` i a continuació es calculen les *abilities* per a cada equip. En la Taula 3.8 es troben els valors d'aquest paràmetre per a cada un dels equips:

Taula 3.8. Valors de les *abilities* segons el model de *Bradley-Terry* per a cada un dels equips

Posició	Equip	*Ability*	Posició	Equip	*Ability*
1	AS Roma	9.991	11	Torino FC	1.078
2	Juventus	8.945	12	AC Milan	1
3	SSC Napoli	3.209	13	ACF Fiorentina	0.907
4	Sassuolo Calcio	2.703	14	Sampdoria	0.796
5	Inter	1.928	15	US Palermo	0.755
6	Atalanta	1.663	16	Udinese Calcio	0.649
7	Genoa CFC	1.537	17	Empoli FC	0.614

8	Lazio Roma	1.477	18	Bologna FC	0.612
9	Carpi FC	1.349	19	Hallas Verona	0.563
10	Chievo Verona	1.214	20	Frosinone Calcio	0.412

Un cop es tenen els valors d'*ability* per a tots els equips és fàcil calcular la probabilitat entre dos equips. Per exemple, la probabilitat que la Juventus guanya a l'ACF Fiorentina és:

$$\mathrm{P}(\text{Juventus guanya Fiorentina}) = \frac{\alpha_{Juventus}}{\alpha_{Juventus} + \alpha_{Fiorentina}} = \frac{8.945}{8.945 + 0.907} = 0.908$$

És a dir, la probabilitat de que la Juventus guanyi a la Fiorentina és d'un 90.8%.

Es decideix de manera arbitrària que en funció de les probabilitats obtingudes es consideraran els següents resultats per al partit:

- P(i guanya j) $\in$ [0, 0.4] $\rightarrow$ Guanya j (visitant)
- P(i guanya j) $\in$ (0.4, 0.6] $\rightarrow$ Empat
- P(i guanya j) $\in$ (0.6, 1] $\rightarrow$ Guanya i (local)

Conclusions (2)

Després d'haver realitzat l'anàlisi amb aquest nou model es comprova la seva capacitat predictiva amb el Boxplot de la Figura 3.3:

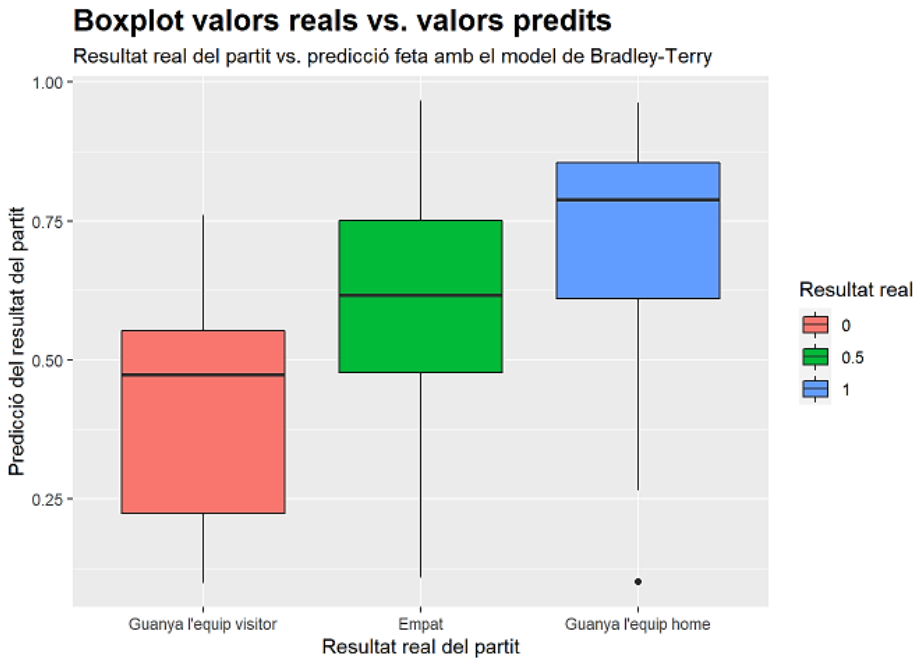

Figura 3.3. Boxplot dels resultats reals dels partits vs. la predicció del resultat amb *Bradley-Terry*

En aquest boxplot s'ha graficat en l'eix de les abscisses el resultat real del partit (guanya l'equip visitant (*visitor*), empat o guanya l'equip local (*home*)). En l'eix de les ordenades es representen les probabilitats predites de que l'equip local guanyi.

S'observa que el model prediu força bé, doncs el boxplot vermell, el qual fa referència al fet que l'equip visitant guanya es troba representat en l'interval [0.20, 0.60] aproximadament. És a dir, quan el resultat real del partit és que guanya l'equip visitant, la predicció realitzada amb el model de *Bradley-Terry* diu que guanyarà l'equip visitant o el resultat del partit serà un empat, però en cap cas guanya l'equip local (el boxplot es troba representat en l'interval [0.20, 0.55] aproximadament pel que fa a l'eix de les ordenades). Pel que fa al segon boxplot, la predicció és força acurada quan el resultat real del partit és un empat. Finalment, l'últim boxplot conclou que quan el partit és guanyat per l'equip local, la predicció prediu el mateix, ja que la caixa blava es troba representada en l'interval [0.62, 0.78] aproximadament.

S'arriba a la mateixa conclusió amb la següent matriu de confusió de la Taula 3.9:

Taula 3.9. Matriu de confusió, valors reals vs. valors predits per al model de *Bradley-Terry*

		Valors reals		
		Guanya l'equip local	Empat	Guanya l'equip visitant
Predicció	Guanya l'equip local	49	14	7
	Empat	14	10	14
	Guanya l'equip visitant	3	4	13

A partir de la taula també es calcula el percentatge d'encert:

$$\%\ \text{encert} = \frac{49 + 10 + 13}{128} = 0{,}5625 \rightarrow 56.25\%$$

Així doncs, en un 56.25% dels casos el model de *Bradley-Terry* plantejat prediu bé el resultat del partit. Tenint en compte que si es volgués predir de manera totalment aleatòria el resultat d'un partit la probabilitat d'encert seria d'un 33.33%, amb el model de *Bradley-Terry* aconseguim un 23% més de precisió.

Per acabar l'anàlisi del model realitzat s'investiga com es va desenvolupar la temporada 2015/2016 de la lliga italiana *Serie A* per comprovar si els resultats obtinguts són similars als reals.

L'edició 2015/2016 de la *Serie A* va tenir un inici imprevist, amb els següents equips lluitant per aconseguir la *Scudetto:* el Napoli, la Juventus, l'Inter, la Fiorentina i la Roma. El primer equip comptava amb la presència de l'argentí Gonzalo Higuaín, el qual posseïa en aquell moment el títol de màxim golejador en aquesta lliga. La Juventus va tenir un mal començament de temporada, doncs després dels 10 primers partits només havia aconseguit sumar 10 punts, col·locant-se en la dotzena posició en la classificació. Pel que fa a l'Inter, el seu objectiu era classificar-se per poder jugar a l'*Europa League*, però els seus resultats el situaven fins i tot en tercera posició (el model de B-T el col·loca en cinquena posició, i finalment l'equip s'acaba classificant quart). L'equip amb seu a Florencia es situava en quarta posició, liderant la lliga en possessió mitjana i precisió de pas. La Roma, l'únic equip que havia sigut capaç de preocupar a la Juventus en les últimes temporades, era considerat l'equip preferit per guanyar el campionat (també ho considera el model de B-T, assignant-li l'*ability* més gran de tots els equips), però una crisi va sorgir, fent que s'acomiadés l'entrenador García i es substituís per Lucas Spalletti (Serie A 2015-16: Half Season Review - StatsBomb | Data Champions, n.d.). L'evolució de la competició va ser estranya, no obstant això va acabar de la manera més previsible: amb la Juventus aconseguint la cinquena *Scudetto* consecutiva. Després d'un mal començament, l'equip d'Allegri va començar una ratxa sense precedents, guanyant 25 de 26 jornades, deixant enrere només 2 punts després d'un empat. Així doncs, la Juventus es va fer amb el campionat a falta de dues jornades per al final (Serie A 2015/2016: Final Review - StatsBomb | Data Champions, n.d.).

Per veure si coincideixen els resultats obtinguts amb el model de *Bradley-Terry* i els resultats finals de la classificació (*Classifica* | *Lega Serie A*, n.d.) es mostra a la Taula 3.10:

Taula 3.10. Classificació dels equips de la lliga *Serie A* i les seves corresponents *abilities* segons el model de *Bradley-Terry*

Equip	**Posició segons classificació temporada 2015/2016**	**Punts**	**Posició segons *ability calculada a partir del model B-T***	***ability***
Juventus	1	91	2	8.945
SSC Napoli	2	82	3	3.209
AS Roma	3	80	1	9.991
Inter	4	67	5	1.928

ACF Fiorentina	5	64	13	0.907
Sassuolo Calcio	6	61	4	2.703
AC Milan	7	57	12	1
Lazio Roma	8	54	8	1.477
Chievo Verona	9	50	10	1.214
Empoli FC	10	46	17	0.614
Genoa CFC	11	46	7	1.537
Torino FC	12	45	11	1.078
Atalanta	13	45	6	1.663
Bologna FC	14	42	18	0.612
Sampdoria	15	40	14	0.796
US Palermo	16	39	15	0.755
Udinese Calcio	17	39	16	0.649
Capri FC	18	38	9	1.349
Frosinone Calcio	19	31	20	0.412
Hellas Verona	20	28	19	0.563

Les caselles marcades de color vermell corresponen a aquells equips que han rebut una classificació segons el model de *Bradley-Terry* com a mínim 4 posicions inferior a la classificació real. Per contra, les caselles de color verd indiquen que l'equip ha rebut una classificació segons B-T com a mínim 4 posicions superior a la classificació real.

La Taula 3.10 mostra que la classificació final per a la temporada 2015/2016 difereix de les *abilities* obtingudes a partir del model de *Bradley-Terry* i això pot ser degut a 2 motius. Primer, que el model no és complert en el sentit que pot haver-hi factors que condicionen el resultat (lesions, ratxes, fitxatges a meitat de temporada,...) que no s'han tingut en compte. Per altra banda, la divisió de les mostres d'entrenament i de validació s'ha fet segons un criteri cronològic implicant que els equips que varen estar més en forma a l'inici de la lliga tinguessin *abilities* més altes.

Destaca per sobre de la resta el posicionament de l'ACF Fiorentina, ja que en la classificació oficial va quedar dins dels 5 primers i el model de *Bradley-Terry* col·loca a

aquest equip en dotzena posició. A la Taula 3.11 s'analitza el comportament d'aquest equip durant les últimes 12 jornades (només les dotze últimes jornades, ja que són les que corresponen a la part *test* de la base de dades, amb la que s'ha realitzat el model) i es pot trobar una explicació al perquè d'aquesta mala classificació amb el model de *Bradley-Terry*.

Taula 3.11. Classificació de l'ACF Fiorentina en les últimes 12 jornades de la temporada 2015/2016 de la lliga italiana *Serie A*

Jornada	Equip local	Equip visitant	Resultat
27	ACF Fiorentina	SSC Napoli	1-1
28	AS Roma	ACF Fiorentina	4-1
29	ACF Fiorentina	Hellas Verona	1-1
30	Frosinone Calcio	ACF Fiorentina	0-0
31	ACF Fiorentina	Sampdoria	1-1
32	Empoli FC	ACF Fiorentina	2-0
33	ACF Fiorentina	Sassuolo Calcio	3-1
34	Udinese Calcio	ACF Fiorentina	2-1
35	ACF Fiorentina	Juventus	1-2
36	Chievo Verona	ACF Fiorentina	0-0
37	ACF Fiorentina	US Palermo	0-0
38	Lazio Roma	ACF Fiorentina	2-4

Els resultats per a l'ACF Fiorentina en les últimes 12 jornades no van ser molt favorables. Només va aconseguir dues victòries en les jornades 33 i 38, en la resta va perdre o empatar contra el contrincant. Sembla que aquests últims partits de la temporada no són representatius per a aquest equip, sobre el qual ja s'ha comentat que es col·locava en quarta posició a la meitat de la temporada i va acabar en cinquena posició. Així doncs, és possible que si la partició de les mostres *train* i *test* s'hagués seleccionat de diferent manera els resultats haguessin sigut uns altres.

Després d'aplicar el PPDAC per segon cop s'ha arribat a un model que és capaç de donar quina és la probabilitat de què un equip guanyi un partit. La funció `BTm` proporciona els paràmetres, anomenats *abilities*, necessaris per poder calcular cada una d'aquestes probabilitats en funció dels equips que s'enfrontin.

3.1.3. Aplicació instrument (*checklist*)

Un cop analitzat el primer estudi de cas seguint el cicle del PPDAC s'aplica la taula *checklist* desenvolupada i descrita en la secció 2.2 del present treball (veure taula 3.12):

Taula 3.12. Aplicació instrument *checklist* per al primer estudi de cas

Pregunta	Resposta	Comentaris
PROBLEM		
Comprensió i definició del problema		
El problema s'ajusta amb el coneixement actual?	Sí	El problema de la predicció en l'esport és actual i no està resolt.
Està clar com es farà per respondre a la pregunta?	Sí	S'avaluarà la capacitat predictiva de models en mostres de validació.
Està clara la pregunta real que es vol fer per afrontar el problema?	Sí	Quin és el percentatge d'encert en resultats de partits de futbol?
Es segueix algun marc o estratègia de formulació de preguntes (e.g. PICO, PECO, PICOT, SPIDER, Other)?	No	
Es coneix l'objectiu i el tipus de la pregunta?	Sí	És un estudi predictiu.
S'ha discutit i pensat en possibles qüestions com ara les implicacions ètiques dels possibles resultats?	No procedeix	Són dades obertes i no hi ha cap qüestió ètica.
S'ha pensat en el tipus de problemes que es podran trobar en aquest estudi?	Sí	S'estarà limitat per no poder avarcar moltes lligues i pel fet que les prediccions en els esports d'equip tenen molta variabilitat.
PLAN		
Què mesurar i com?		
Es té clar quines característiques es mesuraran?	Sí	Està clara la variable resposta basada en el resultat del partit.
Disseny de l'estudi		
S'ha descrit clarament l'objectiu de la recerca?	Sí	Anticipar el resultat d'un partit de futbol.
S'ha descrit clarament el disseny de l'estudi?	Sí	Abans de començar la recollida de dades s'ha fet un pla de recollida i d'anàlisi.
S'ha descrit clarament la població de l'estudi?	Sí	La població són els equips de la lliga italiana *Serie A*.

Es té clar si l'estudi és exploratori o confirmatori?	Sí	És exploratori ja que no es vol confirmar cap hipòtesi.
S'és conscient dels problemes del disseny de l'estudi?	Sí	Per exemple, els resultats no seran extrapolables a altres lligues.
Recollida de dades ("Collecting")		
Es té clar el procés de recollida de les dades?	Sí	Les dades s'obtenen d'un paquet de R.
Es coneixen les variables que es recolliran?	Sí	Són variables resum habituals dels partits de futbol.
Es coneixen les variables resposta que es recolliran?	Sí	Sí. Resultat (Guanyar/Empatar/Perdre) i número de gols de cada equip (Local/Visitant).
Es coneixen les variables d'exposició que es recolliran?	No procedeix	
Es recolliran variables potencialment confusores o modificadores?	No	No es tindrà accés a més variables de les que hi ha recollides al paquet. Lesions, sancions, fitxatges i altres no es recolliran.
Es coneixen les dificultats logístiques que implica la recollida de les dades?	Sí	Al estar en un paquet de R, es preveuen poques dificultats.
Es tindrà en compte la variabilitat de les dades en la decisió?	Sí	S'és coneixedor de la incertesa dels resultats d'esdeveniments esportius.
Passos-cicle i pensament computacional		
Es té suficient coneixement i maneig del programa de maneig de dades?	Sí	Es farà amb el paquet R que es conegut.
S'han tingut en compte apartats del cicle computacional davant la ciència de dades?	Sí	S'ha tingut en compte el cicle *tidyverse*.
S'han tingut en compte aspectes de FAIR amb les dades?	Sí	En aquest cas, estan disponibles al paquet de R.
Enregistrament de les dades ("Recording")		
Es coneixen quines són les unitats d'observació?	Sí	La unitat d'observació és el partit.
Es tenen identificades les fonts d'on provenen les dades?	Sí	Són descrites al paquet `engsoccerdata`, per exemple: https://github.com/footballcsv/en-england, http://en.wikipedia.org/wiki/The_Football_League, http://www.rsssf.com/engpaul/fla, http://www.espn.co.uk/football, http://www.statto.com, http://www.11vs11.com, http://www.worldfootball.net
S'és conscient del volum de dades que s'hauran de treballar?	Sí	El conjunt de dades complert té 25404 registres i 8 columnes que és un volum manegable per R.

DATA		
Supervisar/Auditar les dades		
Són les dades fiables?	Sí	Les pàgines web d'on s'han obtingut les dades són fiables.
Maneig de les dades		
Es té un pla pel maneig de les dades?	No procedeix	Les dades estaran en la memòria de R i no es guardaran enlloc.
Es té algun sistema per mantenir la privacitat de les dades?	No procedeix	Les dades són públiques.
Netejar les dades ("Cleaning")		
Es té un pla pel tractament de les possibles inconsistències de les dades?	Sí	Malgrat ser dades depurades s'ha comprovat la consistència de les dates i del número de gols.
ANALYSIS		
Classificació de les dades ("Sort of data")		
Són les dades reals?	Sí	Són dades extretes de la lliga italiana *Serie A*.
Són les dades simulades?	No	
Es tenen clars els principis bàsics de les dades?	Sí	Malgrat que hi ha principis per reduir els errors, aquestes dades ja estan depurades.
Tècniques estadístiques utilitzades		
La descripció dels mètodes estadístics és correcte i està completada?	Sí	S'ha fet descriptiva de les dades.
Es tenen en compte les premisses que es faran servir durant les anàlisis?	Sí	En tots els models s'assumeixen certes premisses que es tenen en compte.
S'ha fet un ajust o valoració correcta de possibles variables confusores?	No	No es té accés a altres variables.
S'ha fet un tractament de les dades faltants o *missings*?	No procedeix	
S'ha realitzat anàlisi multivariable?	Sí	S'han ajustat models multivariables.
S'ha realitzat anàlisi multivariant?	No	Una alternativa era considerar els gols a casa i fora com dues respostes a analitzar conjuntament (Poisson bivariada), però no s'ha realitzat.
Construcció de taules i gràfics		
S'han emprat eines gràfiques per resumir els resultats?	Sí	S'han fet gràfics per exemple per avaluar la capacitat predictiva.
Has emprat taules per resumir els resultats?	Sí	Per exemple, la matriu de confusió.
S'han tingut en compte els resultats dels gràfics o de les taules per detectar problemes en les dades?	Sí	Per exemple, el sorprenent desajust entre l'observat i l'esperat per l'equip ACF Fiorentina.
S'ha seguit una guia EQUATOR per als anàlisis?	Sí	S'ha seguit la guia TRIPOD per predicció.

CONCLUSIONS		
Interpretació		
Es connecten els resultats (gràfic, taula, resum dades) amb el coneixement existent del problema?	Sí	S'ha fet un resum de la temporada 2015 de la *Serie A* connectant amb els resultats obtinguts.
Tenen sentit els resultats tenint en compte el coneixement actual?	Sí	S'anticipava que la capacitat predictiva no seria molt alta, com així ha estat.
S'han evitat anàlisis selectius o p-hacking?	Sí	Només s'han fet 2 anàlisis i el segon ha estat condicionat per una limitació (manca d'independència en els gols d'un mateix partit) del primer.
S'han interpretat els resultats basant-se en mesures d'efecte?	Sí	Les *abilities* proporcionades indiquen la capacitat de l'equip.
S'ha distingit correctament la causalitat d'associació de la correlació?	Sí	Al ser un estudi observacional, no es poden fer relacions de causa-efecte.
S'ha sabut diferenciar la rellevància de l'efecte amb la significació estadística?	Sí	S'ha evitat parlar de significació estadística.
Conclusions		
S'és conscient de les limitacions a les conclusions basades en com es van mesurar i/o recopilar les dades?	Sí	Falten variables que podien haver fet augmentar la capacitat predictiva: lesions, sancions, fitxatges,...
Les conclusions depenen de les fonts de les dades i els seus anàlisis?	No procedeix	
Comunicació		
És correcta la comunicació de la descripció de les dades?	Sí	S'ha comunicat amb suficient detall com per fer reproduïble l'estudi.
La comunicació en aquest estudi el fa reproduïble o replicable?	Sí	S'explica de manera detallada tot l'anàlisi i es pot accedir al codi a través del repositori Github.
És acceptable la comunicació de la presentació de les figures i/o taules?	Sí	S'han presentat taules i figures segons els estàndards per cada tipus de resultat.
Es comunica la incertesa sobre fets, números i ciència?	No	
Es té clar a qui es comunica?	Sí	A un públic divers interessat en l'estadística esportiva (des de managers, entrenadors, jugadors, fins a investigadors de les ciències del camp de l'esport).
Impacte		
Hi ha un impacte polític?	No	
Hi ha un impacte econòmic?	Sí	Un major coneixement del resultats amb antelació pot condicionar els premis de les cases d'apostes, per exemple.
Hi ha un impacte social?	No	
S'ha tingut en compte l'output acadèmic?	Sí	Es pot derivar un article científic d'aquest treball.

3.2. Estudi de cas 2

3.2.1. Resum executiu

Per a aquest estudi es disposa d'un joc de dades amb informació de 3962 jugadors retirats de bàsquet de l'NBA des de la creació d'aquesta competició fins el 31/07/2019. Per a cada un d'aquests jugadors es disposa de diverses variables, com per exemple el nom del jugador, la seva ètnia o altres característiques físiques com poden ser l'alçada o el pes.

En la taula 3.13 es mostra un resum bàsic de la temàtica de les dades i dels anàlisis que es duran a terme:

Taula 3.13. Resum de les dades de l'estudi del cas 2

Disseny de l'estudi	Temàtica	Mètode estadístic
Longitudinal	Mortalitat en els jugadors de bàsquet de la NBA	Anàlisi de supervivència Riscos Competitius

Origen de les dades

Les dades que es faran servir per a aquest anàlisi són les mateixes que les emprades en l'article de Martinez et al. publicat a l'any 2019 i s'han extret del supplemental material d'aquest article i el seu repostori actualitzat que està ubicat a Github (Casals, 2022). Tal i com expliquen els autors en l'article, les dades van ser obtingudes a partir de combinar la informació de diverses pàgines web a causa de incongruències en algunes de les variables recollides.

Antecedents

Practicar esport de manera freqüent és beneficiós per a la salut física i mental de les persones (Malm et al., 2019). Tot i així, estudis recents mostren que el risc de morir per algunes malalties concretes podria ser superior en jugadors de lligues professionals que en la població en general (Harmon et al., 2015; Morales et al., 2022).

Objectiu

L'objectiu d'aquest estudi és identificar els factors que afecten la longevitat/esperança de vida tenint en compte diferents causes de mort en una població de jugadors ja retirats de la lliga americana de bàsquet (NBA).

3.2.2. Resolució del cas seguint el PPDAC

Problem

El risc de mortalitat en esportistes d'elit amb bases de dades grans o lligues professionals no ha estat molt estudiat fins al moment, excepte algun estudi per a la lliga americana de futbol NFL o per a l'NBA però amb metodologies poc acurades.

Com s'ha comentat, un estudi recent compara el risc de mortalitat en jugadors de l'NBA amb la població general (Martínez et al., 2019). L'article conclou que el risc de mortalitat en els jugadors de bàsquet està relacionat amb l'ètnia i l'alçada.

A partir de les mateixes dades en el present anàlisi es vol identificar si hi ha factors que influeixen en el temps de supervivència dels jugadors de bàsquet en funció de la seva causa de mort.

Plan

Per resoldre el problema plantejat s'aplicaran models d'**anàlisi de supervivència** per explorar el temps que transcorre fins a un esdeveniment. En aquest cas, però caldrà aplicar **mètodes de riscos competitius**, ja que l'esdeveniment d'interès serà la mort per diferents causes.

A l'hora de fer un anàlisi de supervivència amb riscos competitius es poden seguir dues filosofies diferents:

- Modelitzar les causes específiques de risc (*cause-specific hazards*)
- Modelitzar la funció d'incidència acumulada (*cumulative incidence function*)

En la primera filosofia cada un dels riscos és analitzat individualment i es censuren els individus que han mort per altres causes. Aquest enfoc és adequat quan es vol conèixer quins factors estan associats a cada un dels riscos. Per contra, la segona filosofia

s'utilitza per determinar factors associats a la incidència acumulada d'una causa concreta, en aquest cas sense censurar als individus que han mort per altres motius.

Tenint en compte el problema plantejat l'estratègia que més s'adequa és la primera, ja que es pot estudiar cada una de les causes de mort per separat. Un dels models més emprats dins aquesta filosofia és el **model de riscos proporcionals de Cox** (*Cox's proportional hazards model*), també anomenat regressió de Cox (*Cox's regression*). Aquesta metodologia estableix un model per a cada una de les *j* causes específiques de mort, definit com:

$$\lambda_j(t|\boldsymbol{Z}) = \lambda_j(t|\boldsymbol{Z_0})\, e^{\boldsymbol{\beta_j' Z}}, \qquad j = 1, \dots, k$$

on,

- j representa cadascuna de les causes de mort.
- t representa el temps de supervivència. Ens recorda que el risc pot variar al llarg del temps.
- β_j és un vector de dimensió $p \times 1$ (sent *p* el nombre de coeficients del model) corresponents a les covariables i factors del model per cada causa de mort. Els coeficients mesuren l'impacte de cada variable x_i.
- Z és un vector de dimensió $p \times 1$ amb totes les covariables.
- $\lambda_j(t|\boldsymbol{Z_0})$ s'anomena risc de referència (*baseline hazard*). Correspon al valor del risc per la causa j si el valor de totes les covariables fossin igual a 0 i els factors categòrics prenguessin la categoria de referencia. $\lambda_j(t|\boldsymbol{Z})$ és el risc associat a un individu amb les característiques Z.

Les quantitats e^{β_j} s'anomenen *hazard ratios* (HR), que no és res més que el rati de les taxes de risc instantànies entre dos grups d'individus. En aquest cas, al emprar el model de Cox s'assumeix que aquest *hazard rati* és constant durant tot l'estudi. Un valor de β_j superior a 0, o equivalentment un HR superior a 1 indica que a mida que el valor de la i-èssima covariable augmenta, el risc de l'esdeveniment augmenta i, en conseqüència, el temps de supervivència disminueix. Per contra, un valor de β_j inferior a 0, o equivalentment un HR inferior a 1 indica que la covariable i-èssima està positivament associada amb la probabilitat de l'esdeveniment, i relacionada negativament amb el temps de supervivència.

Es podria resumir de la següent manera:

Per a les variables numèriques:

- HR = 1 → No hi ha efecte entre la covariable i el temps de supervivència
- HR < 1 → Hi ha una reducció en el risc per valors alts de la covariable
- HR > 1 → Hi ha un augment en el risc per valors alts de la covariable

Per a les variables factor:

- HR = 1 → No hi ha efecte entre el factor i el temps de supervivència
- HR < 1 → Respecte a la categoria de referència, hi ha una reducció del risc
- HR > 1 → Respecte a la categoria de referència, hi ha un augment del risc

És important remarcar que per poder aplicar aquest model s'ha de realitzar una assumpció, i aquesta és que cada un dels riscos és independent. Dit d'una altra manera, aquells individus censurats (aquells que no han patit l'esdeveniment) per una causa de mort *X* concreta haguessin tingut el mateix risc de mort per *X* independentment del seu desenllaç final (*Competing Risk Analysis | Columbia Public Health*, n.d.; *Cox Proportional-Hazards Model - Easy Guides - Wiki - STHDA*, n.d.; Porta et al., 2007). A més a més, un altre premissa intrínseca del model de Cox és la proporcionalitat de riscos al llarg del temps.

Data

Com ja s'ha comentat anteriorment les dades són obtingudes del repositori de l'article *Mortality of NBA Players: Risk Factors and Comparison with the General US Population* (Martínez et al., 2019) i contenen la informació de 3962 jugadors ja retirats de l'NBA des de la seva creació el 1946 fins al 2019. Les variables que formen la base de dades es mostren a la Taula 3.14:

Taula 3.14. Descripció de la base de dades de l'estudi de cas 2

	Variable	Etiquetes/rang	Descripció
1	id	1 – 4375	Identificador del jugador
2	player		Nom del jugador
3	birthdate	30/01/1902 – 08/11/1999	Data de naixement del jugador
4	pos	C, C-F, F, F-C, F-G, G, G-F	Posició ocupada pel jugador en els partits
5	place	USA/No USA	Lloc de naixement del jugador

6	etni	Black, White, Duda, Mezcla	Raça del jugador
7	debut	30/01/1902 – 31/06/2019	Data en que el jugador va debutar a la lliga
8	from	1946 – 2018	Any en que el jugador va debutar a la lliga
9	fins	1947 – 2020	Any en que el jugador va deixar de jugar a la lliga
10	dateevent	19/11/1957 – 31/07/2019	Data en que va succeir l'esdeveniment. Per als jugadors que no han mort es va fixar com a data el 31/07/2019
11	cens	0, 1	Indica si el jugador està viu o mort en el moment de la recollida de dades. Pren el valor 0 per als jugadors morts i el valor 1 per als jugadors vius
12	death	Yes, No	Indica si el jugador està viu o mort en el moment de la recollida de dades
13	ageevent	19 – 99	Edat del jugador en el moment de l'esdeveniment
14	ageleft	17.45 – 45.75	Edat en que es comença a jugar a la NBA
15	ageright	19.73 – 99.55	Edat en que es mor o es deixa de seguir
16	kilos	60.33 – 163	Pes del jugador en quilograms
17	cms	160.02 – 231.14	Alçada del jugador en centímetres
18	cms2	160.02 – 231.14	Alçada del jugador en centímetres
19	bmi	16.98 – 31.91	Índex de massa corporal del jugador
20	games	1 – 1611	Nombre de partits jugats
21	lefthanded	Yes, No	Indica si el jugador és esquerrà

22	ICD-10	Categories segons la *International Classification of Diseases*	Causa de mort del jugador
23	Validación Death Cause	Causa de mort segons els investigadors	Causa de mort del jugador
24	Codificación final	Causa de mort segons els metges de l'estudi	Causa de mort del jugador
25	Cause_death	Accident/Homicide/Suicide, Cancer, Cardiac Disease, Natural/Old age, Other, Unknown	Causa de mort del jugador

Les categories en que es va dividir la variable posició (*pos)* que ocupa el jugador són les següents:

- C → Center
- C-F → Center-Forward
- F → Forward
- F-C → Forward-Center
- F-G → Forward-Guard
- G → Guard
- G-F → Guard-Forward

Les variables *IDC-10*, *Valicación Death Cause* i *Codificación final* van ser afegides per diversos investigadors clínics que van col·laborar amb els autors de l'article publicat per Martínez et al a l'any 2019. Per simplificar l'anàlisi s'ha creat la variable *Cause_death* i s'han agrupat totes les causes de mort en 6 categories, basant-se en l'article *Vital statistics and early death predictors of North American professional basketball players: A historical examination* (Lemez et al., 2018):

- Accident/Homicide/Suicide
- Cancer
- Cardiac Disease
- Natural/Old age
- Other
- Unknown

Els individus amb causa de mort desconeguda (*Unknown*) no s'analitzaran, ja que en aquesta categoria es concentren una gran varietat de causes de mort desconegudes, fent que no tingui sentit analitzar-les.

Analysis

Per dur a terme l'anàlisi de supervivència es carreguen les dades a R i es fa servir la funció `coxph` del paquet `survival` per crear cada un dels models, corresponents a cada una de les possibles causes de mort. Les variables que s'han decidit incloure en el model són l'ètnia del jugador, l'edat del jugador en finalitzar la carrera a l'NBA, l'alçada i l'any de l'última temporada que l'esportista va jugar a l'NBA. L'elecció de les variables ha estat l'esmentada per poder comparar els resultats amb els obtinguts en l'article de Martínez et al. publicat l'any 2019. A la Taula 3.15 es mostra l'estimació puntual i per interval dels *hazard ratios* per a cada una de les variables en funció del model.

Taula 3.15. Estimació puntual i per Interval dels *hazard ratios* per al model general i per a cada possible causa de mort

Model 0 (Martínez et al., 2019)	**HR**	$\boldsymbol{IC_{95\%}(HR)}$
Ètnia (Blanc vs. Afroamericà)	0.71	[0.57, 0.88]
Edat al final de la carrera a la NBA	1.10	[1.07, 1.12]
Alçada	1.02	[1.01, 1.03]
Any de l'última temporada a la NBA	0.98	[0.97, 0.99]
Model 1 – Accident/Homicidi/Suïcidi	**HR**	$\boldsymbol{IC_{95\%}(HR)}$
Ètnia (Blanc vs. Afroamericà)	0.63	[0.33, 1.19]
Edat al final de la carrera a la NBA	0.98	[0.92, 1.05]
Alçada	1.03	[1.00, 1.07]
Any de l'última temporada a la NBA	0.98	[0.96, 1.00]
Model 2 – Càncer	**HR**	$\boldsymbol{IC_{95\%}(HR)}$
Ètnia (Blanc vs. Afroamericà)	0.75	[0.49, 1.15]
Edat al final de la carrera a la NBA	1.10	[1.06, 1.15]
Alçada	1.02	[1.00, 1.05]
Any de l'última temporada a la NBA	0.99	[0.97, 1.01]

Model 3 – Malaltia cardiovascular	**HR**	$\boldsymbol{IC_{95\%}(HR)}$
Ètnia (Blanc vs. Afroamericà)	0.60	[0.40, 0.92]
Edat al final de la carrera a la NBA	1.10	[1.06, 1.15]
Alçada	1.02	[1.00, 1.05]
Any de l'última temporada a la NBA	0.99	[0.97, 1.01]
Model 4 – Natural/Vellesa	**HR**	$\boldsymbol{IC_{95\%}(HR)}$
Ètnia (Blanc vs. Afroamericà)	1.20	[0.30, 4.75]
Edat al final de la carrera a la NBA	1.25	[1.13, 1.39]
Alçada	1.04	[0.99, 1.09]
Any de l'última temporada a la NBA	1.04	[0.96, 1.11]
Model 5 – Altres	**HR**	$\boldsymbol{IC_{95\%}(HR)}$
Ètnia (Blanc vs. Afroamericà)	0.78	[0.43, 1.42]
Edat al final de la carrera a la NBA	1.13	[1.07, 1.20]
Alçada	1.00	[0.97, 1.03]
Any de l'última temporada a la NBA	1.01	[0.99, 1.04]

Com ja s'ha explicat en l'apartat *Plan* aquelles variables amb un *hazard ratio* igual a 1 no afecten el temps de supervivència dels individus. Per tant, en els IC del 95% de confiança que inclouen aquest valor no tenim evidència de que el factor o la covariable tingui alguna associació sobre la supervivència d'una causa concreta.

En el model 0, corresponent a l'estudi publicat l'any 2019, totes les variables són rellevants, ja que els seus intervals de confiança per als HR no contenen el valor 1. Els *hazard ratios* d'aquestes variables indiquen que el risc de mortalitat és més baix per a jugadors blancs que per a negres, que com més edat té el jugador en retirar-se més alt és el risc de mortalitat, que els jugadors més alts tenen un risc de mortalitat superior als més baixos i que jugadors que es van retirar fa menys anys tenen un risc de mortalitat més baix que aquells que es van retirar fa més anys.

Per al model 1, el qual fa referència a morts per assassinat, homicidi o suïcidi, totes les variables inclouen el valor 1 dins l'interval de confiança, de manera que és arriscat concloure que alguna d'elles té realment un efecte sobre el temps de supervivència. Tot

i així les variables alçada i any de l'última temporada a la NBA inclouen el valor 1 en un dels seus límits, de manera que sent menys estrictes es podria dir que l'alçada influeix incrementant el risc de mortalitat en aquest tipus de mort i que els jugadors retirats fa més temps tenen un risc de mortalitat inferior.

En el segon model es tracten les morts causades per càncer. En aquest cas només la variable que fa referència a l'edat que té el jugador en el moment de retirar-se és rellevant per conèixer el temps de supervivència. El seu valor del HR és superior a 1, de manera que com més avançada és l'edat del jugador quan es retira més risc hi ha de que mori per càncer. Tal i com passava en el model 1 la variable alçada conté el valor 1 en un dels seus intervals, de manera que essent més flexible també es podria concloure que com més alt és un jugador més risc hi ha de que mori per càncer.

El tercer model és per a les causes de mort causades per malalties cardiovasculars. En aquest cas l'ètnia del jugador està associada amb el temps de supervivència del jugador, fent que els jugadors blancs tinguin un risc de mortalitat inferior als jugadors afroamericans. L'edat en què el jugador es retira també es significativa, augmentant el risc de mortalitat com més elevada és l'edat. A més a més en aquest model l'interval de confiança de l'alçada torna a tenir l'1 com a límit inferior, fent que aquesta afecti de manera negativa al temps de supervivència.

En els últims dos models, referents a morts naturals o causades per vellesa i les morts agrupades en la categoria "altres" només l'edat que té el jugador en el moment de retirar-se és significativa en el risc de mortalitat, fent que com més elevada sigui l'edat més risc de mortalitat hi hagi.

Per visualitzar més fàcilment quines de les variables afecten més a cada una de les causes de mort s'ha fet un *Forest Plot* amb l'ajuda del paquet `ggplot2` de R Cran. En el gràfic es troba marcat amb una línia discontínua el valor 1 i d'aquesta manera es pot veure que les variables que es troben a l'esquerra de la línia discontinua comporten una reducció en el risc i les que es troben a la dreta comporten un augment del risc de mortalitat.

Les variables numèriques s'han reescalat sumant-hi 5 unitats a cada una d'elles per així fer-les més comparables.

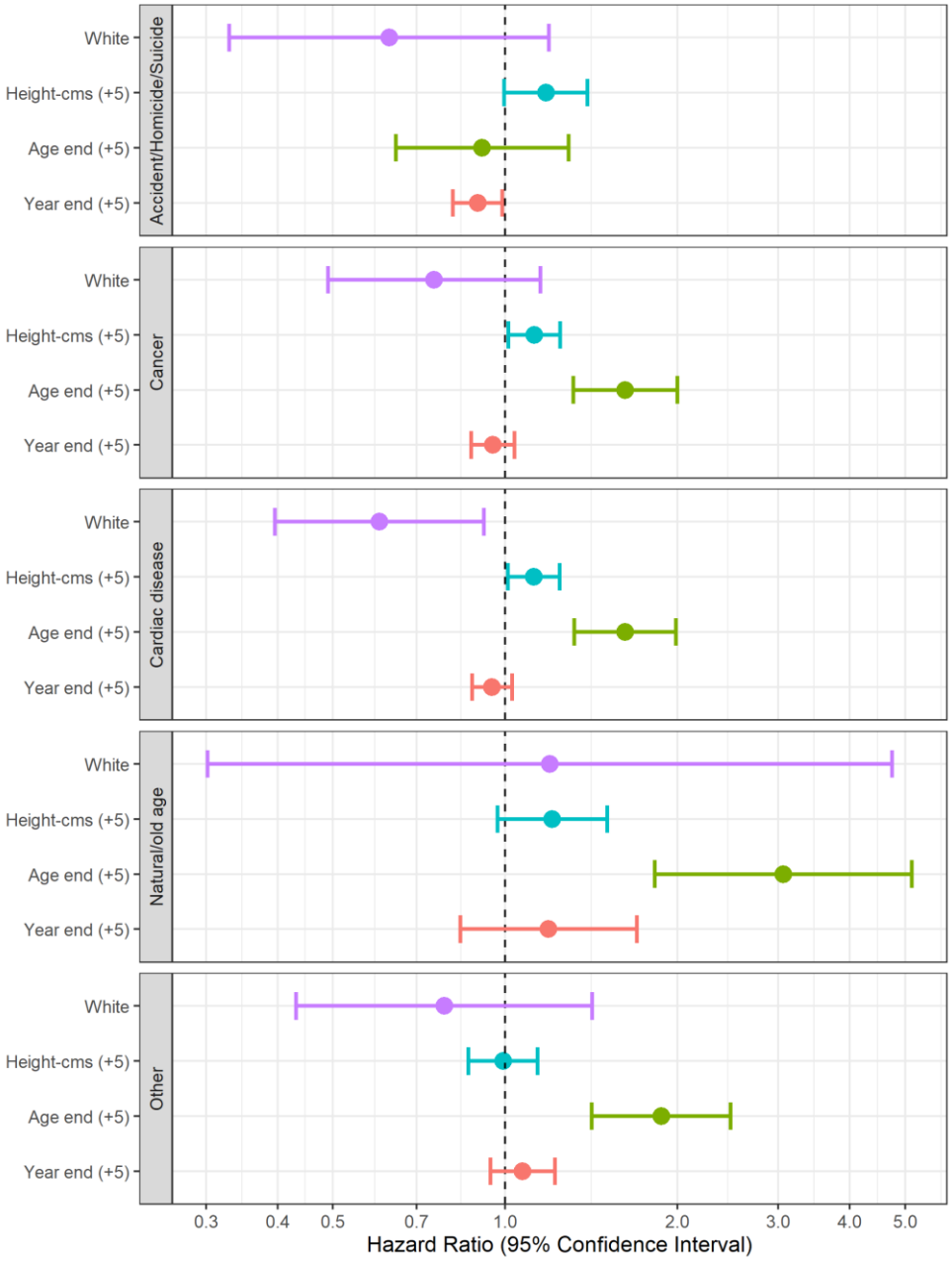

Figura 3.4. *Forest Plot* per als *hazard ratios* per cada una de les causes de mort

Els resultats obtinguts són els mateixos que els comentats anteriorment anteriorment, que es mostren resumits en la següent Taula 3.16:

Taula 3.16. Taula resum amb les variables rellevants per als temps de supervivència en funció de cada causa de mort

Causa de mort	Variables rellevants	Factors que disminueixen el temps de supervivència
Assassinat/ Homicidi/Suïcidi	Alçada	↑ alçada
	Any en retirar-se	↑ any
Càncer	Edat en retirar-se	↑ edat
	Alçada	↑ alçada
Malalties cardiovasculars	Ètnia	Ser d'ètnia afroamericà (respecte a ser de ètnia blanca)
	Edat en retirar-se	↑ edat
	Alçada	↑ alçada
Natural/Vellesa	Edat en retirar-se	↑ edat
Altres	Edat en retirar-se	↑ edat

Conclusions

Observant els resultats dels models específics per a cada una de les causes de mort s'arriba a la conclusió que l'edat al final de la carrera de la NBA és rellevant per a totes les causes de mort excepte per a les causades per assassinat, homicidi o suïcidi, fent que el risc de mortalitat augmenti. Aquest resultat sembla força coherent, doncs per a la població en general és normal que les persones d'edat més avançada pateixin més problemes de salut i en el cas dels jugadors de bàsquet de la NBA els jugadors més vells presenten una càrrega de partits i entrenaments que pot afectar al temps de supervivència. L'alçada també és rellevant en tres dels cinc models plantejats, fent que els jugadors més alts tinguin un risc de mortalitat superior als jugadors baixos amb menys alçada. L'ètnia també és rellevant pel que fa a les morts causades per malalties cardiovasculars. En aquest cas els jugadors blancs tenen un risc de mortalitat inferior als jugadors afroamericans. Finalment l'any en què el jugador és retira també és important a l'hora de mesurar el temps de supervivència d'un individu, fent que aquest augmenti com més recent sigui l'any en el que s'ha retirat.

En el model inicial, el plantejat per Martínez et al. l'any 2019, tots els factors eren rellevants, però sobretot ho eren l'ètnia (els jugadors afroamericans tenien més risc a morir que els blancs) i l'alçada. En aquest anàlisi, mirant cada una de les causes específiques, s'han desgranat una mica més aquestes conclusions i s'ha observat que no és així en les diferents causes de mort, i hi ha altres factors que també poden ser rellevants en funció de la causa específica.

Un cop dut a terme el cicle del PPDAC es creu convenient intentar identificar quina causa específica té una probabilitat de supervivència més elevada. Per fer-ho, es torna a aplicar el cicle:

Problem (2)

Es vol estudiar quina de les causes de mort té un temps de supervivència més elevat.

Plan (2)

Per poder analitzar quina causa de mort té una probabilitat de supervivència més elevada és important conèixer la **funció d'incidència acumulada** $F_j(t)$, que es pot definir com

$$F_j(t) = P(T \leq t, C = j), \qquad j = 1, \dots, k$$

i correspon a la sub-funció de distribució de la probabilitat d'un subjecte a patir l'esdeveniment degut a la causa específica *j*.

Data (2)

Les dades emprades per calcular la funció d'incidència acumulada són les mateixes que les fetes servir anteriorment, sense fer cap modificació.

Analysis (2)

Pel que fa a conèixer quina és la causa de mort amb una probabilitat de supervivència superior es decideix graficar la funció d'incidència acumulada, fent servir les funcions `cuminc` i `ggcompetingrisks` dels paquets `cmprsk` i `survminer`, respectivament.

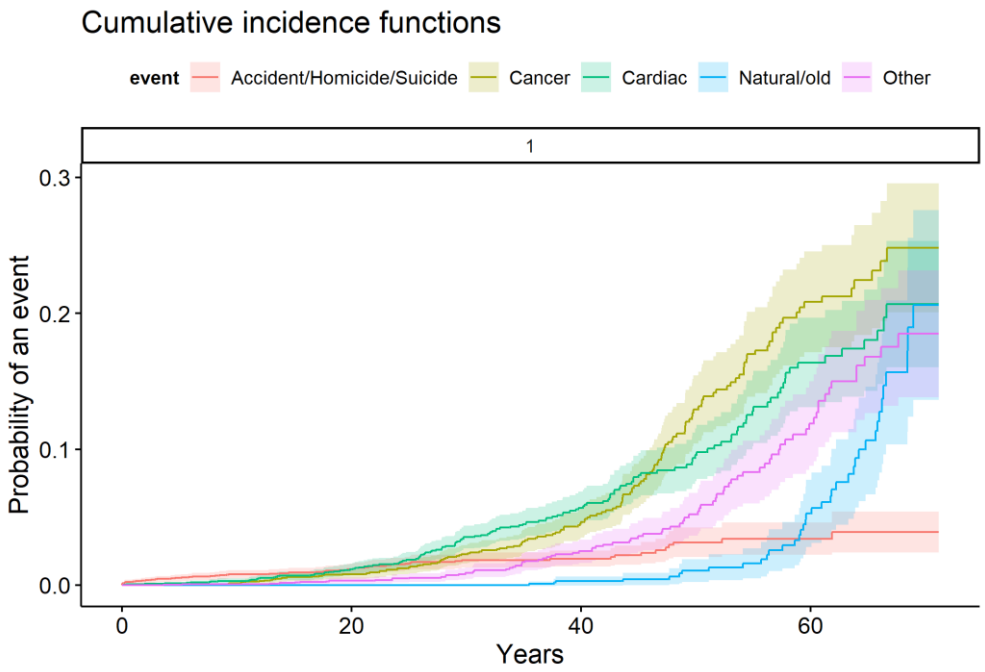

Figura 3.5. *Funció d'incidència acumulada* per a cada una de les causes de mort

S'observa que les causes de mort relacionades amb càncer o amb malalties cardiovasculars tenen una incidència acumulada superior a la resta, tal i com passa amb la població en general (Mathers et al., 2009). Tenint en compte que la funció d'incidència acumulada és complementària a la funció de supervivència s'arriba a la conclusió que aquestes dues causes de mort presenten una probabilitat de supervivència inferior a la resta de causes de mort.

Per a les causes de mort causades per càncer, malalties cardiovasculars, naturals o per vellesa i les morts classificades en la categoria altres la probabilitat de morir comença a augmentar a partir dels 40 anys de deixar de jugar (cal recordar que el nostre esdeveniment és la mort i que el temps de supervivència es calcula des del moment en què el jugador es retira fins que succeeix l'esdeveniment). Per a la causa específica assassinat, homicidi o accident l'evolució és més constant i la probabilitat augmenta abans, a partir dels 20 anys de retirar-se.

És important destacar que al final del seguiment les probabilitats acumulades de totes les causes de morts han de sumar el 100%. A la Taula 3.17 es mostren les freqüències relatives de cada una de les causes de mort:

Taula 3.17. Taula amb les probabilitats acumulades de les causes de mort

Causa de mort	Incidència acumulada de mortalitat
Càncer	31.8%
Malalties cardiovasculars	30.5%
Altres	18%
Assassinat/Homicidi/Suïcidi	11.7%
Natural/Vellesa	8.1%

Així doncs, deixant de banda les causes de mort *Unknown* tal i com s'ha fet a l'estudi, les causes de mort causades per càncer suposen aproximadament un 31.8% del total de morts, seguides molt d'aprop per les defuncions causades per malalties cardiovasculars, les quals representen un 30.5% del total de les morts. Una mica més allunyades es troben les causes de mort categoritzades en "Altres" i assassinat, homicidi o suïcidi, amb 18 i un 11.7% respectivament. Finalment les causes de mort naturals o degudes a l'edat suposen un 8.1% del total.

Com que la variable edat en el moment de retirar-se és rellevant en quatre dels cinc models i la variable alçada també ho és en tres del cinc models, es decideix també fer un gràfic d'incidència acumulada estratificant per cada un d'aquestes variables:

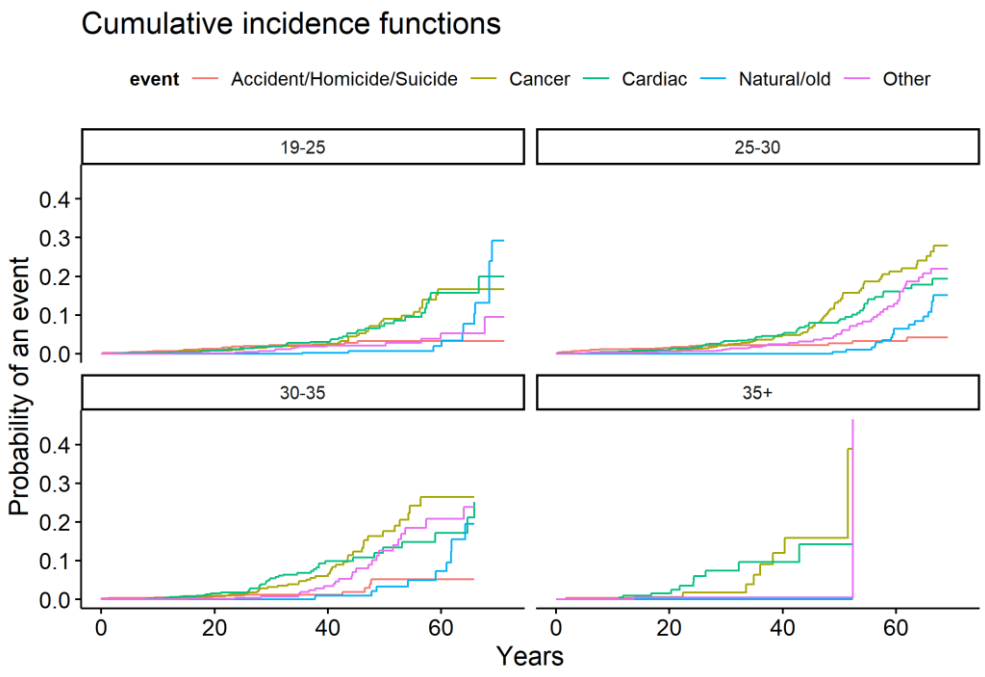

Figura 3.6. *Funció d'incidència acumulada* per cada causa de mort en funció de l'edat en retirar-se

La variable que fa referencia a l'edat del jugador en el moment de retirar-se era rellevant en tots els models excepte el que tractava les causes de mort per assassinats, homicidis o suïcidis. Es demostra que la probabilitat de que succeeixi l'esdeveniment per càncer, malalties cardiovasculars, naturalesa o per altres motius augmenta amb l'edat. A més a més, tal i com s'ha observat en el gràfic anterior, el temps de supervivència quan l'edat de retirada és inferior a 35 anys en general supera els 60 anys, mentre que quan l'edat en retirar-se supera els 35 aquest temps de supervivència és clarament inferior a 60. Sembla una conclusió bastant lògica, doncs és normal que com més edat tingui el jugador menys temps de vida li quedi, independentment de la causa per la que finalment mori.

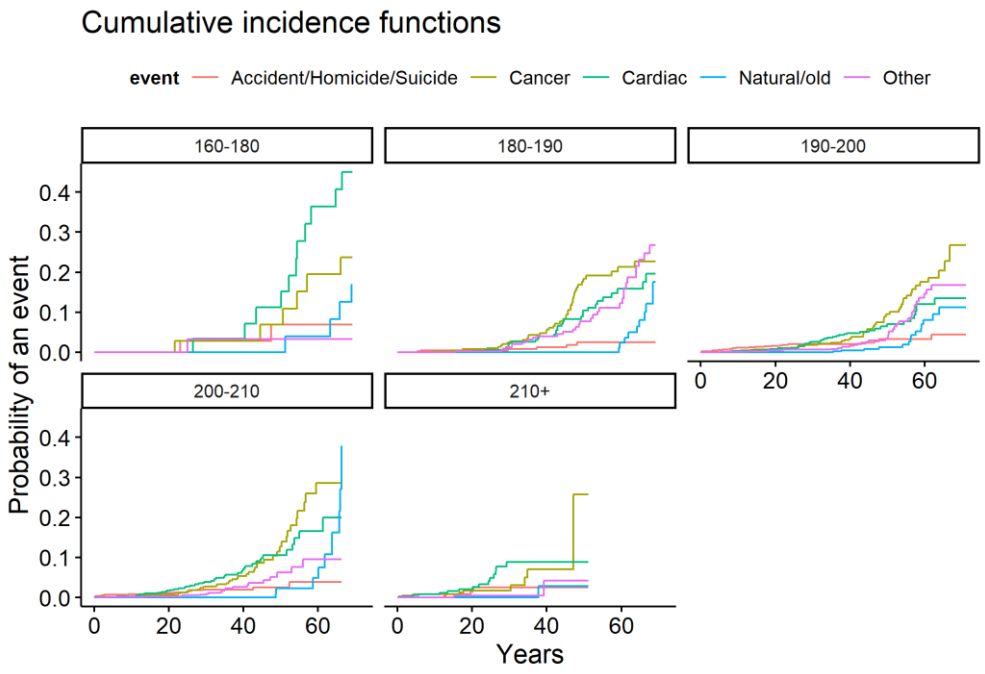

Figura 3.7. *Funció d'incidència acumulada* per cada causa de mort en funció de l'alçada

S'observa que els jugadors amb un alçada superior als 210cm tenen un temps de supervivència clarament inferior a la resta de jugadors, doncs el temps màxim entre el moment de retirar-se i l'esdeveniment (que recordem que és la mort) és aproximadament de 50 anys, mentre que per a la resta d'alçades aquest temps supera els 60 anys.

Cal comentar que per poder realitzar amos gràfics s'ha hagut de realitzar categories de les dues variables i l'elecció dels punts de tall són arbitraris.

Tot i que la variable relacionada amb l'ètnia només és rellevant en un dels models també es farà un gràfic d'incidència acumulada en funció de les ètnies, doncs al ser una variable categòrica és més fàcil realitzar-lo i també pot proporcionar informació interessant:

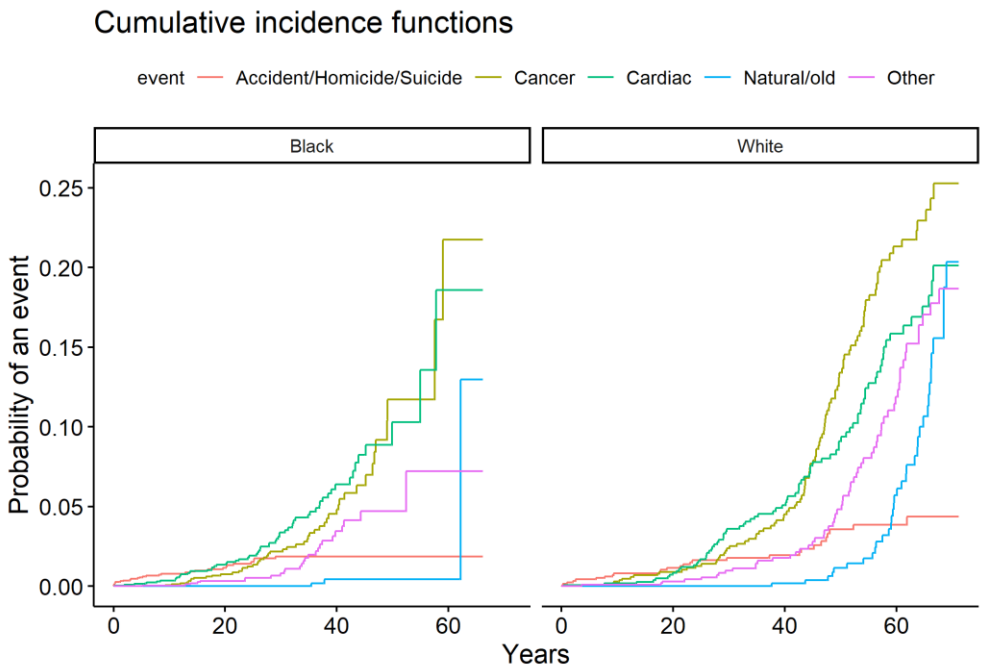

Figura 3.8. *Funció d'incidència acumulada* per a cada causa de mort en funció de l'ètnia

S'observa que la primera causa de mort tant en jugadors blancs com negres és el càncer. Pel que fa als jugadors de raça afroamericana la segona causa de mort és deguda a malalties cardiovasculars, mentre que per als esportistes blancs ho són les morts naturals o per vellesa, seguides molt de prop per les morts cardíaques i les englobades dins la categoria altes. Per contra, per als jugadors negres la incidència acumulada per a les morts naturals o categoritzades com altres es troben molt lluny de les dues primeres causes de mort. En ambdues races les morts per accidents, homicidis o suïcidis són minoritàries. S'observa també que el temps de supervivència tant per als jugadors blancs com per als negres supera els 60 anys, però si es mira més en detall el gràfic es veu clarament que el temps de supervivència per als jugadors blancs és lleugerament superior al dels jugadors afroamericans.

Conclusions (2)

Gràcies a les representacions gràfiques i al càlcul de les incidències acumulades realitzades s'observa amb més claredat quins són els factors que afecten cada una de

les causes de mort, tractades amb més detall en l'apartat de conclusions realitzat en el primer cicle del PPDAC aplicat a aquest joc de dades.

A més a més, s'ha afegit l'anàlisi de quina de les causes de mort té una probabilitat de supervivència inferior, i es conclou que les causes de mort que presenten un risc més elevat són les degudes a càncers o malalties cardiovasculars, seguit d'altres malalties, accidents, homicidis o suïcidis i finalment morts naturals o deguda a la vellesa dels jugadors de bàsquet.

En un treball futur es planteja incloure els jugadors no retirats dins l'anàlisi tenint en compte el truncament per l'esquerra.

3.2.3. Aplicació instrument (*checklist*)

Un cop analitzat el segon estudi de cas seguint el cicle del PPDAC s'aplica la taula *cheklist* desenvolupada i descrita en la secció 2.2 del present treball (veure Taula 3.18):

Taula 3.18. Aplicació instrument *checklist* per al segon estudi de cas

Pregunta	Resposta	Comentaris
PROBLEM		
Comprensió i definició del problema		
El problema s'ajusta amb el coneixement actual?	Sí	Es coneix la importància d'estudiar la salut dels atletes.
Està clar com es farà per respondre a la pregunta?	Sí	S'ha dissenyat un procés tenint en compte un pla de recollida i anàlisi de dades complert.
Està clara la pregunta real que es vol fer per afrontar el problema?	Sí	Indicar si hi ha factors que influeixen en el temps de supervivència dels jugadors de l'NBA retirats en funció de la seva causa de mort.
Es segueix algun marc o estratègia de formulació de preguntes (e.g. PICO, PECO, PICOT, SPIDER, Other)?	No	
Es coneix l'objectiu i el tipus de la pregunta?	Sí	És un estudi predictiu.
S'ha discutit i pensat en possibles qüestions com ara les implicacions ètiques dels possibles resultats?	No procedeix	Les dades provenen de fonts públiques.
S'ha pensat en el tipus de problemes que es podran trobar en aquest estudi?	Sí	Per exemple que algunes causes de mort siguin inespecífiques.

PLAN		
Què mesurar i com?		
Es té clar quines característiques es mesuraran?	Sí	Les variables de l'article d'en Martínez et al (2019).
Disseny de l'estudi		
S'ha descrit clarament l'objectiu de la recerca?	Sí	Identificar el temps de supervivència i els factors que influeixen en les diferents causes de mort dels jugadors ja retirats de l'NBA des de la seva fracció fins l'any 2019.
S'ha descrit clarament el disseny de l'estudi?	Sí	Es tracta d'un estudi de cohort retrospectiu.
S'ha descrit clarament la població de l'estudi?	Sí	En l'article de Martínez et al (2019).
Es té clar si l'estudi és exploratori o confirmatori?	Sí	És exploratori ja que no es vol confirmar cap hipòtesi de les diferents causes de mort.
S'és conscient dels problemes del disseny de l'estudi?	Sí	Poden haver problemes en obtenir tota la informació de manera 100% fiable al obtenir les dades de diferents fonts.
Recollida de dades ("Collecting")		
Es té clar el procés de recollida de les dades?	Sí	S'ha utilitzat la base de dades de Martínez et al (2019). Es van registrar totes les dades de diferents fonts d'internet, especialment les pàgines web de *Sports Reference* i Newspapers.com
Es coneixen les variables que es recolliran?	Sí	Les mateixes que l'article de Martínez et al (2019) afegint la variable causa de mort.
Es coneixen les variables resposta que es recolliran?	Sí	Temps fins a les diferents causes de mort.
Es coneixen les variables d'exposició que es recolliran?	Sí	Sí, les descrites a la Taula 3.14.
Es recolliran variables potencialment confusores o modificadores?	Sí	Edat al final carrera NBA i any de la darrera temporada a la NBA.
Es coneixen les dificultats logístiques que implica la recollida de les dades?	Sí	Van ser descrites en l'article de Martínez et al (2019), i ara amb la variable causa de mort encara s'ha estat més conscient al treballar amb moltes fonts per reportar aquesta variable.
Es tindrà en compte la variabilitat de les dades en la decisió?	Sí	Es reporten els IC al 95% per exemple en els *hazard ratis*.
Passos-cicle i pensament computacional		
Es té suficient coneixement i maneig del programa de maneig de dades?	Sí	Funcions de *wrangling data* i *scraping* han estat prèviament ja utilitzades a l'article de Martínez et al (2019).

S'han tingut en compte apartats del cicle computacional davant la ciència de dades?	Sí	Dins del cicle *tidyverse* sí.
S'han tingut en compte aspectes de FAIR amb les dades?	No	
Enregistrament de les dades ("Recording")		
Es coneixen quines són les unitats d'observació?	Sí	Cadascun dels jugadors de l'NBA.
Es tenen identificades les fonts d'on provenen les dades?	Sí	Són descrites a l'article de Martinez et al (2019); per exemple: https://www.newspapers.com, https://www. basketball-reference.com, https://www.nba.com, https://www.fiba.com, https://www.apbr.org, https://www.nytimes.com, https://www.latimes.com, https://www.espn.com
S'és conscient del volum de dades que s'hauran de treballar?	Sí	Serà una base de dades amb 3962 jugadors retirats de l'NBA i 25 variables.
DATA		
Supervisar/Auditar les dades		
Són les dades fiables?	Sí	En l'article de Martínez et al (2019) es descriu acuradament el procés de validació de les dades, tot i que no es podria assegurar 100% la fiabilitat en cadascuna de les característiques.
Maneig de les dades		
Es té un pla pel maneig de les dades?	SÍ	Descrit a l'article Martínez et al (2019).
Es té algun sistema per mantenir la privacitat de les dades?	No	
Netejar les dades ("Cleaning")		
Es té un pla pel tractament de les possibles inconsistències de les dades?	Sí	És descrit a l'article Martínez et al (2019). Per exemple s'han realitzat algunes imputacions.
ANALYSIS		
Classificació de les dades ("Sort of data")		
Són les dades reals?	Sí	Dades de jugadors de l'NBA retirats desde la creació de la competició fins al 2019.
Són les dades simulades?	No	
Es tenen clars els principis bàsics de les dades?	Sí	La construcció de les dades es feta a partir de l'article Martínez et al (2019).
Tècniques estadístiques utilitzades		
La descripció dels mètodes estadístics és correcte i està completada?	Sí	Per exemple s'han presentat dues filosofies dels mètodes de riscos competitius.

Es tenen en compte les premisses que es faran servir durant les anàlisis?	Sí	Per exemple el risc de proporcionalitat en el model de Cox, o les premisses d'independència o dependència.
S'ha fet un ajust o valoració correcta de possibles variables confusores?	No	Edat al final carrera NBA i any de la darrera temporada a l'NBA.
S'ha fet un tractament de les dades faltants o missings?	No procedeix	
S'ha realitzat anàlisi multivariable?	Sí	Resum del model de supervivència de l'article de Martínez et al (2019).
S'ha realitzat anàlisi multivariant?	No	Diferents causes de mort són diferents variables resposta.
Construcció de taules i gràfics		
S'han emprat eines gràfiques per resumir els resultats?	Sí	Figures per mostrar per exemple les incidències acumulades per les diferents causes de mort.
Has emprat taules per resumir els resultats?	Sí	Mesures resum dels factors rellevants en el temps de supervivència segons les diferents causes de mort.
S'han tingut en compte els resultats dels gràfics o de les taules per detectar problemes en les dades?	Sí	Per exemple veure si en el *forest plot* els intervals i l'estimació puntual tenien sentit segons si la categoria de referència en la variable ètnia era una o altra.
S'ha seguit una guia EQUATOR per als anàlisis?	Sí	Per exemple s'han reportat intervals de confiança segons la guia STROBE.
CONCLUSIONS		
Interpretació		
Es connecten els resultats (gràfic, taula, resum dades) amb el coneixement existent del problema?	Sí	Gràcies a aquests tres elements es pot contestar el problema inicial plantejat.
Tenen sentit els resultats tenint en compte el coneixement actual?	Sí	Per exemple les causes de mort cardiovasculars i de càncer són les més freqüents, igual que passa en la població general.
S'han evitat anàlisis selectius o p-hacking?	Sí	Es mostren només mesures de l'efecte i no s'han fet anàlisis en la cerca de resultats significatius.
S'han interpretat els resultats basant-se en mesures d'efecte?	Sí	S'ha reportat l'interval de confiança al 95%.
S'ha distingit correctament la causalitat d'associació de la correlació?	Sí	És un estudi observacional on l'objectiu és merament descriptiu i exploratori.
S'ha sabut diferenciar la rellevància de l'efecte amb la significació estadística?	Sí	S'ha descrit una taula resum dels resultats on justament es parla només de rellevància.
Conclusions		
S'és conscient de les limitacions a les conclusions basades en com es van mesurar i/o recopilar les dades?	Sí	Les causes de mort desconegudes. S'ha descrit el seu tractament en l'estudi.

Les conclusions depenen de les fonts de les dades i els seus anàlisis?	No	
Comunicació		
És correcta la comunicació de la descripció de les dades?	Sí	S'ha comunicat amb suficient detall com per fer reproductible l'estudi.
La comunicació en aquest estudi el fa reproduïble o replicable?	Sí	S'explica de manera detallada tot l'anàlisi i es pot accedir al codi a través del repositori Github.
És acceptable la comunicació de la presentació de les figures i/o taules?	Sí	S'han presentat taules i figures segons els estàndards per cada tipus de resultat.
Es comunica la incertesa sobre fets, números i ciència?	Sí	Per exemple amb intervals de confiança.
Es té clar a qui es comunica?	Sí	A un públic divers interessat en l'estadística esportiva (des de managers, entrenadors, jugadors, fins a investigadors de les ciències del camp de l'esport).
Impacte		
Hi ha un impacte polític?	No	
Hi ha un impacte econòmic?	Sí	Un major coneixement de les causes de mort dels jugadors pot ajudar a adaptar la carrera professional de cada jugador i que els equips en puguin treure un rendiment econòmic
Hi ha un impacte social?	No	Les dades poden ajudar sobretot a científics, a l'associació de jugadors de l'NBA i els mateixos actors per impulsar conjuntament una estratègia sanitària basada en sistemes de vigilància amb millors registres de dades.
S'ha tingut en compte l'output acadèmic?	Sí	S'ha utilitzat un article científic de l'any 2019, i es té pensat publicar l'estudi de diferents causes de mort en el futur.

4. Conclusions

El cicle del PPDAC i l'instrument creat en aquest treball dota d'eines i recursos als investigadors per poder donar resposta a una qüestió científica d'una manera més organitzada i metòdica. Pensar què es vol resoldre (*Problem*) és essencial per saber com es durà a terme l'anàlisi (*Plan*), ja que aquest es pot enfocar de diferent manera en funció de la pregunta plantejada. Aquest segon pas està estretament lligat amb el següent, les dades (*Data*): és important conèixer les característiques de les dades i les seves limitacions. Un bon treball en els dos passos previs pot fer que el següent pas (*Analysis*) sigui menys feixuc. L'últim pas (*Conclusions*) és tant o més important que les anteriors ja que és la fase en què es tradueixen els resultats numèrics o gràfics a un llenguatge apropiat pel públic al qual s'adreça l'estudi.

L'instrument *checklist* creat permetrà no només basar-se de manera subjectiva dels diferents apartats del cicle PPDAC sinó verificar quines preguntes no s'han tingut en compte per completar correctament el cicle i resoldre el problema de la millor manera possible. De fet, sense l'ús de l'instrument en els dos exemples exposats s'ha acabat utilitzant la metodologia de forma cíclica, repetint algunes de les fases i permetent així una millora constant en la investigació de l'estudi.

La importància d'aquest instrument ve donada en dos sentits. Per una banda, fa palesa la rellevància de l'estadístic en les fases prèvies a l'anàlisi. Tradicionalment, s'ha vist als estadístics com uns professionals que coneixien un ampli ventall d'eines numèriques per realitzar l'anàlisi d'unes dades. La funció d'aquesta professió va molt més enllà i el rol de l'estadístic s'ha d'ampliar en la fase de disseny d'un estudi on pot alertar de potencials problemes posteriors i, per tant, ajudar a configurar millor el protocol de treball. Per un altre banda, aquest instrument més que donar solucions als investigadors, busca despertar el pensament estadístic i plantejar-se qüestions que poden ser cabdals per un bon desenllaç de l'estudi. El fet de que es tracti d'un cicle i no d'un procés lineal s'ajusta més a la realitat del mètode científic de prova i error; no obstant, un bon treball d'inici hauria de reduir el nombre d'iteracions que es realitzen dins del cicle.

Som conscients que aquest instrument té limitacions i, malgrat que, totes les qüestions estan basades en la literatura científica, s'ha de treballar en un futur consultant amb els altres experts sobre les qüestions realment rellevants i útils per acabar tenint una eina que faciliti el flux de treball al llarg d'un estudi científic.

5. Bibliografia

Albert, J., Glickman, M. E., Swartz, T. B., & Koning, R. H. (2017). Handbook of statistical methods and analyses in sports. In Handbook of Statistical Methods and Analyses in Sports. doi: 10.1201/9781315166070

Statistical thinking in sports, (2007).

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., & Witmer, J. (2005). GAISE college report. In American Statistical Association. *Retrieved October*, *16*.

AlMuhayfith, F. E., Alzaid, A. A., & Omair, M. A. (2016). On bivariate Poisson regression models. *Journal of King Saud University - Science*, *28*(2), 178–189. doi: 10.1016/J.JKSUS.2015.09.003

Arnold, P., & Franklin, C. (2021). What Makes a Good Statistical Question? *Journal of Statistics and Data Science Education*, *29*(1). doi: 10.1080/26939169.2021.1877582

Booth, A., Noyes, J., Flemming, K., Moore, G., Tunçalp, Ö., & Shakibazadeh, E. (2019). Formulating questions to explore complex interventions within qualitative evidence synthesis. *BMJ Global Health*, *4*. doi: 10.1136/bmjgh-2018-001107

Broman, K. W., & Woo, K. H. (2018). Data Organization in Spreadsheets. *American Statistician*, *72*(1). doi: 10.1080/00031305.2017.1375989

Bullock, G. S., Hughes, T., Arundale, A. H., Ward, P., Collins, G. S., & Kluzek, S. (2022). Black Box Prediction Methods in Sports Medicine Deserve a Red Card for Reckless Practice: A Change of Tactics is Needed to Advance Athlete Care. *Sports Medicine*. doi: 10.1007/s40279-022-01655-6

Büttner, F., Ardern, C. L., Blazey, P., Dastouri, S., McKay, H. A., Moher, D., & Khan, K. M. (2021). Counting publications and citations is not just irrelevant: It is an incentive that subverts the impact of clinical research. In British Journal of Sports Medicine (Vol. 55, Issue 12). doi: 10.1136/bjsports-2020-103146

Casals, M. (2022). *marticasals/Reproducibility_Mortality_NBA_Players: NBA Data*. doi: 10.5281/ZENODO.6469668

Cervone, D., Amour, A. D. ', Bornn, L., & Goldsberry, K. (2016). *A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes*. doi: 10.1080/01621459.2016.1141685

Classifica | Lega Serie A. (n.d.). Retrieved from https://www.legaseriea.it/it/serie-a/classifica/2015-16

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015a). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *European Urology, and Journal of Clinical Epidemiology*, *162*, 55–63. doi: 10.7326/M14-0697

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015b). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *European Urology*, *67*(6). doi: 10.1016/j.eururo.2014.11.025

Competing Risk Analysis | Columbia Public Health. (n.d.). Retrieved from https://www.publichealth.columbia.edu/research/population-health-methods/competing-risk-analysis

Cox, D. R., Kartsonaki, C., & Keogh, R. H. (2018). Big data: Some statistical issues. *Statistics and Probability Letters*, *136*. doi: 10.1016/j.spl.2018.02.015

Cox Proportional-Hazards Model - Easy Guides - Wiki - STHDA. (n.d.). Retrieved from http://www.sthda.com/english/wiki/cox-proportional-hazards-model

Davies, K. S. (2011). Evidence Based Library and Information Practice Commentary Formulating the Evidence Based Practice Question: A Review of the Frameworks. *Evidence Based Library and Information Practice*, *6*(2).

Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software*, *12*, 1–12. doi: 10.18637/jss.v012.i01

GAISE College Report ASA Revision Committee. (2016). Guidelines for assessment and instruction in statistics education (GAISE) college report 2016. In Report.

Goldschmidt, G., & Matthews, B. (2022). Formulating design research questions: A framework. *Design Studies*, *78*. doi: 10.1016/j.destud.2021.101062

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650). doi: 10.1136/bmj.39489.470347.ad

Harmon, K. G., Asif, I. M., Maleszewski, J. J., Owens, D. S., Prutkin, J. M., Salerno, J. C., Zigman, M. L., Ellenbogen, R., Rao, A. L., Ackerman, M. J., & Drezner, J. A. (2015). Incidence, cause, and comparative frequency of sudden cardiac death in national collegiate athletic association athletes a decade in review. *Circulation*, *132*(1). doi: 10.1161/CIRCULATIONAHA.115.015431

Hayes Davenport, T., & Patil, D. (2012). *Data Scientist: The Sexiest Job of the 21st Century*. Retrieved from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Hidalgo, B., & Goodman, M. (2013). Multivariate or multivariable regression? In American Journal of Public Health (Vol. 103, Issue 1). doi: 10.2105/AJPH.2012.300897

Jeffrey O., B., Briggs, W., & Triola, M. (2003). *Statistical reasoning for everyday life. Boston, MA: Addison-Wesley.* (2nd ed.). Boston.

Kingsman, T. (2016). *Extended Bradley-Terry Models*.

Lang, T. A., & Altman, D. G. (2015). Basic statistical reporting for articles published in Biomedical Journals: The “Statistical analyses and methods in the published literature” or the SAMPL guidelines. In International Journal of Nursing Studies (Vol. 52, Issue 1). doi: 10.1016/j.ijnurstu.2014.09.006

Leek, J. T., & Peng, R. D. (2015). What is the question? Mistaking the type of question being considered is the most common error in data analysis. In Science (Vol. 347, Issue 6228). doi: 10.1126/science.aaa6146

Lemez, S., Wattie, N., Lawler, T., & Baker, J. (2018). Vital statistics and early death predictors of North American professional basketball players: A historical examination. *Journal of Sports Sciences*, *36*(14). doi: 10.1080/02640414.2017.1409607

Ley, C., Wiele, T. van de, & Eetvelde, H. van. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, *19*(1). doi: 10.1177/1471082X18817650

Lopez, M. J., & Matthews, G. (2015). Skidmore College Creative Matter Building an NCAA Men' s Basketball Predictive Model and Quantifying Its Success Recommended Citation "Building an NCAA men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports*, 11–12. Retrieved from https://creativematter.skidmore.edu/math_fac_schol.https://www.degruyter.com/view/j/jqas.2015.11.issue-1/jqas-2014-0058/jqas-2014-0058.xml?format=INT

Lopez, M. J., Matthews, G. J., & Baumer, B. S. (2018). How often does the best team win? A unified approach to understanding randomness in north american sport. *Annals of Applied Statistics*, *12*(4). doi: 10.1214/18-AOAS1165

Lopez, M., & Skidmore College. (n.d.). *Lab 11: Paired comparison models*.

Macdonald, B. (2020). Recreating the Game: Using Player Tracking Data to Analyze Dynamics in Basketball and Football. *Harvard Data Science Review*, *2*(4). doi: 10.1162/99608f92.6e25c7ee

MacKay, R. J., & Oldford, R. W. (2000). Scientific method, statistical method and the speed of light. *Statistical Science*, *15*(3). doi: 10.1214/ss/1009212817

Malm, C., Jakobsson, J., & Isaksson, A. (2019). *Physical Activity and Sports-Real Health Benefits: A Review with Insight into the Public Health of Sweden*. doi: 10.3390/sports7050127

Mansournia, M. A., Collins, G. S., Nielsen, R. O., Nazemipour, M., Jewell, N. P., Altman, D. G., & Campbell, M. J. (2021). A CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): Explanation and elaboration. In British Journal of Sports Medicine (Vol. 55, Issue 18). doi: 10.1136/bjsports-2020-103652

Martínez, J. A., Langohr, K., Felipo, J., & Casals, M. (2019). Mortality of NBA players: Risk factors and comparison with the general US population. *Applied Sciences (Switzerland)*, *9*(3). doi: 10.3390/app9030500

Mathers, C. D., Boerma, T., Fat, D. M., & Mathers, C. (2009). Global and regional causes of death. *British Medical Bulletin*, *92*, 7–32. doi: 10.1093/bmb/ldp028

McNamara, A., Horton, N. J., & Baumer, B. S. (2017). Greater Data Science at Baccalaureate Institutions. In Journal of Computational and Graphical Statistics (Vol. 26, Issue 4). doi: 10.1080/10618600.2017.1386568

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2012). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery*, *10*(1), 28–55. doi: 10.1016/J.IJSU.2011.10.001

Morales, J. S., Valenzuela, P. L., Saco-Ledo, G., Castillo-García, A., Carabias, C. S., McCrory, P., Santos-Lozano, A., & Lucia, A. (2022). Mortality Risk from Neurodegenerative Disease in Sports Associated with Repetitive Head Impacts: Preliminary Findings from a Systematic Review and Meta-Analysis. In Sports Medicine (Vol. 52, Issue 4). doi: 10.1007/s40279-021-01580-0

Nielsen, R. O., Simonsen, N. S., Casals, M., Stamatakis, E., & Mansournia, M. A. (2020). Methods matter and the “too much, too soon” theory (part 2): What is the goal of your sports injury research? Are you describing, predicting or drawing a causal inference? In British Journal of Sports Medicine. doi: 10.1136/bjsports-2020-102144

P. Curley, James. (2016). *Engsoccerdata: English Soccer Data 1871-2106. R Package Version 0.1, 5.*

Porta, N., Gómez, G., Calle, M. L., Malats, N., & Porta, U. (2007). *COMPETING RISKS METHODS*. Retrieved from http://www-eio.upc.es/%7Enporta/.correspondingauthor:N´

Praharaj, S. K., & Ameen, S. (2020). How to choose research topic? *Kerala Journal of Psychiatry*, *33*(1). doi: 10.30834/kjp.33.1.2020.188

Sabbag, A., Garfield, J., & Zieffler, A. (2018). Assessing statistical literacy and statistical reasoning: The reali instrument. *Statistics Education Research Journal*, *17*(2). doi: 10.52041/serj.v17i2.163

Sainani, K., & Chamari, K. (2022). Wish List for Improving the Quality of Statistics in Sport Science. *International Journal of Sports Physiology and Performance*. doi: 10.1123/ijspp.2022-0023

Sainani, K. L., Borg, D. N., Caldwell, A. R., Butson, M. L., Tenan, M. S., Vickers, A. J., Vigotsky, A. D., Warmenhoven, J., Nguyen, R., Lohse, K. R., Knight, E. J., & Bargary, N. (2021). Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. In British Journal of Sports Medicine (Vol. 55, Issue 2). doi: 10.1136/bjsports-2020-102607

Schield, M. (2017). *Statistical literacy*.

Serie A 2015-16: Half Season Review - StatsBomb | Data Champions. (n.d.). Retrieved from https://statsbomb.com/articles/soccer/serie-a-2015-16-half-season-review/

Serie A 2015/2016: Final Review - StatsBomb | Data Champions. (n.d.). Retrieved from https://statsbomb.com/articles/soccer/serie-a-20152016-final-review/

Skidmore College. (n.d.-a). *Lecture 10: Power rankings*.

Skidmore College. (n.d.-b). *Lecture 11: Statistics in soccer*.

Spiegelhalter, D. (2019). *The art of statistics: Learning from data.* .

Starkings, S., & Albert, J. (2008). Statistical Thinking in Sports, edited by Jim Albert & Ruud H. Koning. *International Statistical Review*, *76*(1). doi: 10.1111/j.1751-5823.2007.00039_10.x

Turner, H., & Firth, D. (2012). Bradley-Terry Models in R: The BradleyTerry2 Package. . *Journal of Statistical Software, 48(9), 1-21.* .

Utts, J. (2014). *Seeing through statistics. Cengage Learning.*

Utts, J. (2021a). Enhancing Data Science Ethics Through Statistical Education and Practice. *International Statistical Review*, *89*(1). doi: 10.1111/insr.12446

Utts, J. (2021b). Statistical Practice Is Not a Spectator Sport. *Harvard Data Science Review*. doi: 10.1162/99608f92.ff65fd7a

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, *6*(5). doi: 10.1098/rsos.181870

Vandenbroucke, J. P., & Pearce, N. (2018). From ideas to studies: How to get ideas and sharpen them into research questions. *Clinical Epidemiology*, *10*. doi: 10.2147/CLEP.S142940

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. In PLoS Biology (Vol. 13, Issue 4). doi: 10.1371/journal.pbio.1002128

Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., Savic, M. D., Garovic, V. D., & Milic, N. M. (2019). Reveal, Don't Conceal: Transforming Data Visualization to Improve Transparency. In Circulation (Vol. 140, Issue 18). doi: 10.1161/CIRCULATIONAHA.118.037777

Whelan, J. T., & Klein, J. E. (2021). BRADLEY–TERRY MODELING WITH MULTIPLE GAME OUTCOMES WITH APPLICATIONS TO COLLEGE HOCKEY. *Mathematics for Applications*, *10*(2). doi: 10.13164/ma.2021.13

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3). doi: 10.1111/j.1751-5823.1999.tb00442.x

Wild, C. J., Utts, J. M., & Horton, N. J. (2018). *What is statistics?. In International handbook of research in statistics education (pp. 5-36). Springer, Cham.* (pp. 5–36).

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1). doi: 10.1111/j.2041-210x.2009.00001.x

6. Annex

Pregunta	Etiqueta	Referència	Explicació
PROBLEM			
Comprensió i definició del problema		(Spiegelhalter, 2019) (Vandenbroucke et al., 2018)	Pensar en el veritable estat del problema ens fa pensar en d'on partim (ex: vaga idea, informació general, preguntes àmplies, preguntes precises, o inclús altres opcions). A més a més, ens pot fer conscient de les nostres necessitats (ex: "Una necessitat d'actuar", "Una necessitat de saber", "Ambdues", Altra)
El problema s'ajusta amb el coneixement actual?	(Sí, No, No està clar, No procedeix)	(Utts, 2021b)	No es pot resoldre un problema sense definir abans què és, i no s'hauria d'intentar resoldre un problema sense qüestionar-te per què ho fas. Cal pensar si el problema s'adapta al coneixement existent.
Està clar com es farà per respondre a la pregunta?	(Sí, No, No està clar, No procedeix)	(Spiegelhalter, 2019)	Abans de començar a cercar, s'ha de tenir clar què s'està cercant i el com.
Està clara la pregunta real que es vol fer per afrontar el problema?	(Sí, No, No està clar, No procedeix)	(GAISE College Report ASA Revision Committee, 2016)	Arribar a preguntes útils que es puguin respondre de manera realista utilitzant dades estadístiques sempre implica molt de pensament i sovint molt treball preparatori. (Wild, C. J., Utts, J. M., & Horton, N. J. (2018))

Es segueix algun marc o estratègia de formulació de preguntes (per exemple, PICO, PECO, PICOT, SPIDER, Other)?	(Sí, No, No està clar, No procedeix)	(Davies, 2011) (Booth et al., 2019) (Praharaj et al., 2020) (Goldschmidt et al., 2022)	Hi ha diversos mètodes que poden ajudar a estructurar la pregunta. Alguns exemples serien: SMART (Sharp/Specific – Measurable – Achievable – Realistic – Timely), PECO (Population/Problem – Exposure – Comparison – Outcome), PICO (Population/Problem – Intervention – Comparison – Outcome), PICOT (Population/Problem – Intervention – Comparison – Outcome – Timeframe), FINER (Feasibility – Interesting – Novel – Ethical – Relevant), SPIDER (Sample – Phenomena of interest – Design – Evaluation – Research type), ECLIPSE (Expectation, Client Group, Location, Impact, Professionals, Service)...
Es coneix l'objectiu i el tipus de la pregunta?	(Sí, No, No està clar, No procedeix)	(Leek et al., 2015) (Nielsen et al., 2020) (MacKay et al., 2000)	Els objectius i tipus de pregunta tenen un lligam. Ens podem trobar amb diferents tipus de preguntes (no és una anàlisi de dades, descriptiu, exploratori, inferencial, predictiu, causal, mecanistic, altra) que caldrà tenir present per marcar adequadament l'objectiu real i apropar-nos a fer millors preguntes.
S'ha discutit i pensat en possibles qüestions com ara les implicacions ètiques dels possibles resultats?	(Sí, No, No està clar, No procedeix)	(Utts, 2021b) (Utts, 2021a)	Per assegurar-se que les anàlisis són adequades cal respondre abans les preguntes d'interès. Per fer-ho cal discutir i raonar les implicacions ètiques dels possibles resultats i com el problema s'adapta al coneixement existent.
S'ha pensat en el tipus de problemes que es podran trobar en aquest estudi?	(Sí, No, No està clar, No procedeix)	(Wild et al., 2018)	L'alfabetització de les dades i estadística intenta tenir present els possibles problemes que hi pot haver (ex: estudis observacionals, confusió, causalitat, problema de les proves múltiples, mida de la mostra

			i significació estadística, reproductibilitat, disminució del risc augmenta realment el risc, risc personalitzat versus risc mitjà, poca intuïció sobre la probabilitat i el risc, ús dels valors esperats per prendre decisions, problema en les enquestes, tipus de biaixos...)
PLAN			
Què mesurar i com?		(Spiegelhalter, 2019)	El pas del "Plan" consisteix en decidir de quines persones/objectes/entitats cal obtenir dades, i de quines coses hauria de "mesurar" i com farem tot això.
Es té clar quines característiques es mesuraran?	(Sí, No, No està clar, No procedeix)	(Wild et al., 2018)	Preguntar-se sobre quines característiques es mesuraran i pensar sobre possibles errors de mesures pot ser important abans d'avançar en aquest cicle. Per exemple, en l'àmbit clínic es coneix la "Clinimetria" que té un conjunt de regles que regeixen l'estructura dels índexs, l'elecció de les variables dels components, l'avaluació de la consistència, la validesa i la capacitat de resposta. La perspectiva clinimètrica proporciona una llar intel·lectual per al judici clínic, la implementació del qual és probable que millori els resultats tant en la investigació clínica com en la pràctica.
Disseny de l'estudi		(Spiegelhalter, 2019)	
S'ha descrit clarament l'objectiu de la recerca?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	Ser explícit sobre l'objectiu de la recerca és un requisit previ en la ciència independentment de la disciplina científica. L'objectiu de l'estudi fa referència a la raó de l'estudi i assenyala la qüestió científica específica que s'està abordant.

S'ha descrit clarament el disseny de l'estudi?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	El disseny de l'estudi fa referència al tipus d'estudi. Alguns exemples de dissenys d'estudis són els RCT i estudis observacionals com ara estudis de cohort, casos i controls o estudis transversals. Com a principi general, el disseny de l'estudi s'ha d'explicar de manera que un altre investigador podria repetir l'estudi exactament.
S'ha descrit clarament la població de l'estudi?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021) (MacKay et al., 2000)	La població d'estudi es refereix a la població d'origen en que es recullen les dades, mentre que la població objectiu es refereix a la població a la qual generalitzarem els resultats de l'estudi; saber la relació entre aquestes dues poblacions és crucial per avaluar la generalització.
Es té clar si l'estudi és exploratori o confirmatori?	(Sí, No, No està clar, No procedeix)	(Nielsen et al., 2020)	Cal tenir present la distinció entre la investigació que busca principalment explorar patrons sense hipòtesis articulades a priori (recerca exploratòria) i la investigació que posa a prova explícitament hipòtesis formulades a priori (investigació confirmatòria).
S'és conscient dels problemes del disseny de l'estudi?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	Ser conscient dels problemes que es deriven del disseny es pensar en la validesa del disseny de l'estudi, la declaració clara i justificació de la mida de la mostra, i les infraccions o "violacions" del disseny.
Recollida de dades ("Collecting")		(Spiegelhalter, 2019)	
Es té clar el procés de recollida de les dades?	(Sí, No, No està clar, No procedeix)	(Utts, 2021a) (Utts, 2021b)	Tenir clar el procés de recollida de dades porta a comprendre conceptes d'estratègies de mostreig (probabilístic/no probabilístic), *web scraping,* recollida de les dades recopilades per l'empresari.

Es coneixen les variables que es recolliran?	(Sí, No, No està clar, No procedeix)	(Arnold et al., 2021)	És necessari preguntar-se quines possibles variables es podrien explorar per respondre la pregunta d'investigació
Es coneixen les variables resposta que es recolliran?	(Sí, No, No està clar, No procedeix)	(MacKay et al., 2000)	Cal preguntar-se si tenim una, varies o cap variable objectiu, resposta o *outcome*.
Es coneixen les variables d'exposició que es recolliran?	(Sí, No, No està clar, No procedeix)	(MacKay et al., 2000)	Hi ha una variable d'interès predominant que representi la variable principal d'exposició/intervenció?
Es recolliran variables potencialment confusores o modificadores?	(Sí, No, No està clar, No procedeix)	(MacKay et al., 2000)	Es té clar quines variables (confusores) poden estar associades amb la variable resposta i amb la variable d'interès principal a la vegada? Es té clar en quins subgrups l'associació de entre l'exposició/intervenció en estudi i la variable resposta pot ser diferent (interacció)?
Es coneixen les dificultats logístiques que implica la recollida de les dades?	(Sí, No, No està clar, No procedeix)	(Arnold et al., 2021)	Cal preguntar-se si es podran recollir dades per respondre aquesta pregunta d'investigació.
Es tindrà en compte la variabilitat de les dades en la decisió?	(Sí, No, No està clar, No procedeix)	(GAISE College Report ASA Revision Committee, 2016)	Cal preguntar-se sobre la variació, variabilitat i tenir-la en compte en les decisions. Emfatitzar el pensament estadístic recau en pensar en l'omnipresència de la variabilitat, i la quantificació i explicació de la variabilitat.
Passos-cicle i pensament computacional			

Es té suficient coneixement i maneig del programa de maneig de dades?	(Sí, No, No està clar, No procedeix)	(Wild et al., 2018) (Horton et al., 2021)	És essencial ser capaç d'abordar un problema i crear, expressar o transformar una solució per a un problema i que aquesta pugui ser desenvolupada per un ordinador.
S'han tingut en compte apartats del cicle computacional davant la ciència de dades?	(Sí, No, No està clar, No procedeix)	(McNamara et al., 2017) (Wickham et al., 2017)	La descripció a nivell computacional s'engloba dins algunes subseccions com: 1. Data Gathering, Preparation, and Exploration 2. Data Representation and Transformation 3. Computing with Data 4. Data Modeling 5. Data Visualization and Presentation 6. Science about Data Science En el mateix sentit per exemple en el programa R s'ha començat a utilitzar molt el paquet *tidyverse* en educació, ja que, permet fomentar el cicle a nivell computacional on té en compte d'altres paquets relacionats amb *data import, visualisation, wrangling, modeling,* i *communication*.
S'han tingut en compte aspectes de FAIR amb les dades?	(Sí, No, No està clar, No procedeix)	(Longo et al., 2016)	Els Principis FAIR ofereixen un conjunt de qualitats precises i mesurables que una publicació de dades hauria de seguir perquè les dades siguin Trobables, Accessibles, Interoperables i Reutilitzables (de l'anglès FAIR – Findable, Accessible, Interoperable, and Reusable).

Enregistrament de les dades ("Recording")		(Spiegelhalter, 2019)	
Es coneixen quines són les unitats d'observació?	(Sí, No, No està clar, No procedeix)	(MacKay et al., 2000)	La població objectiu o "target population" és el col·lectiu d'unitats sobre les quals es vol treure conclusions. Cal tenir cura al especificar tan les unitats com la població objectiu. Per a algunes investigacions pot ser més fàcil definir les unitats o el col·lectiu en termes d'un procés que els genera. En alguns casos, pot ser més convenient fer referència al procés objectiu en lloc de la població objectiu.
Es tenen identificades les fonts d'on provenen les dades?	(Sí, No, No està clar, No procedeix)	(Utts, 2021a) (Utts, 2021b)	És important preguntar-se per la font de les dades per pensar si està ben entesa o revisada segons d'on provingui (*web scraping, random sampling, disseny experimental*,...) per entendre comportaments i conceptes posteriors (consentiment informat, anonimat, aspectes ethics, *dropouts*,...).
S'és conscient del volum de dades que s'hauran de treballar?	(Sí, No, No està clar, No procedeix)	(Cox et al., 2018) (Zuur et al., 2010)	Ser conscient del volum (*small, moderate, big*) ajudarà a preguntar-se si les dades són rellevants per al propòsit de la investigació, per si la qualitat de les dades és adequada pels posteriors anàlisis estadístics. Un dels reptes és de vegades l'excés de confiança en els resultats obtinguts a partir d'anàlisis de grans conjunts de dades, a causa d'una alta precisió però superficial davant estimacions potencialment esbiaixades o a causa d'errors estàndard subestimats. La mida de les dades no elimina la necessitat de disseny adequat de l'estudi i anàlisi estadística.
DATA			

Supervisar/Auditar les dades		(Spiegelhalter, 2019)	
Són les dades fiables?	(Sí, No, No està clar, No procedeix)	(Arnold et al., 2021)	S'ha de fer un esforç per garantir que la mesura retorna valors fiables de les variacions que es pretén mesurar. Això pot implicar la qualitat de les dades, el seu seguiment i la validesa interna. La metrologia, és a dir, les qüestions de mesura, és fonamental per al progrés en molts camps.
Maneig de les dades		(Spiegelhalter, 2019)	
Es té un pla pel maneig de les dades?	(Sí, No, No està clar, No procedeix)	(MacKay et al., 2000)	
Es té algun sistema per mantenir la privacitat de les dades?	(Sí, No, No està clar, No procedeix)	(MacKay et al., 2000)	A l'hora d'emmagatzemar les dades cal pensar en plans de gestió i privadesa de les dades.
Netejar les dades ("Cleaning")		(Spiegelhalter, 2019)	
Es té un pla pel tractament de les possibles inconsistències de les dades?	(Sí, No, No està clar, No procedeix)	(Utts, 2021a)	
ANALYSIS			

Classificació de les dades ("Sort of data")		(Spiegelhalter, 2019)	
Són les dades reals?	(Sí, No, No està clar, No procedeix)	(GAISE College Report ASA Revision Committee, 2016)	Una de les sis recomanacions de la GAISE és justament integrar dades reals amb un context i un propòsit. Això porta sovint a escenaris de la vida real amb múltiples variables.
Són les dades simulades?	(Sí, No, No està clar, No procedeix)	(GAISE College Report ASA Revision Committee, 2016)	Realitzar simulacions pot ajudar a il·lustrar conceptes abstractes.
Es tenen clar els principis bàsics de l'organització de les dades?	(Sí, No, No està clar, No procedeix)	(Broman et al., 2018)	El treball de Broman avisa de recomanacions pràctiques per a organitzar les dades de la fulla de càlcul per a reduir els errors i facilitar les anàlisis posteriors.
Tècniques estadístiques utilitzades		(Spiegelhalter, 2019)	
La descripció dels mètodes estadístics és correcte i està completada?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	Completar correctament els mètodes estadístics emprats és clau per poder replicar o reproduir i entendre els resultats d'un estudi per exemple i la tècnica emprada.
Es tenen en compte les premisses que es faran servir durant les anàlisis?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	De vegades les assumpcions no són explicites i són necessàries per poder entendre tots els procediments o tècniques estadístiques utilitzades.

S'ha fet un ajust o valoració correcta de possibles variables confusores?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	S'empraran tècniques d'anàlisi basades en models per tenir en compte la confusió (principal font de biaix en estudis observacionals/experimentals).
S'ha fet un tractament de les dades faltants o *missing*?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	S'hauran de descriure les dades mancants de les variables recollides i emprar els mètodes més sofisticats (*multiple imputation*) en comptes de mètodes *naive* (anàlisi complert, imputar mitjana, LOCF).
S'ha realitzat anàlisi multivariable?	(Sí, No, No està clar, No procedeix)	(Hidalgo et al., 2013)	El model multivariable s'utilitza per a l'anàlisi amb un *outcome* (variable dependent) i múltiples variables independents (també anomenat predictor o variable explicativa).
S'ha realitzat anàlisi multivariant?	(Sí, No, No està clar, No procedeix)	(Hidalgo et al., 2013)	El model multivariant ("multivariate") s'utilitza per a l'anàlisi amb més d'una variable *outcome* (p. ex. agrupació, PCA, mesures repetides).
Construcció de taules i gràfics		(Spiegelhalter, 2019)	
S'han emprat eines gràfiques per resumir els resultats?	(Sí, No, No està clar, No procedeix)	(MacKay et al., 2000)	La visualització de les dades pot apropar a millorar el pensament estadístic. Segons la cita d'Alberto Cairo del 2015, "*A [data] visualization is any visual display intented to reveal evidence, making the invisible visible*".
S'han emprat taules per resumir els resultats?	(Sí, No, No està clar, No procedeix)	(MacKay et al., 2000)	Les taules requereixen sovint més temps i experiència per aconseguir els mateixos resultats que els gràfics però de vegades segons el què i com es vulgui comunicar a posteriori els resultats cal reconsiderar-ho.

S'han tingut en compte els resultats dels gràfics o de les taules per detectar problemes en les dades?	(Sí, No, No està clar, No procedeix)	(Zuur et al., 2010) (Weissgerber et al., 2015) (Weissgerber et al., 2019)	L'exploració de les dades ajuda a veure possibles perills o amenaçes posteriors en els anàlisis estadístiques. Tenir un protocol establert o seguir guies afins pot ser molt important.
S'ha seguit alguna guia EQUATOR per als anàlisis?	(Sí, No, No està clar, No procedeix)	(Lang et al., 2015) (Collins et al., 2015) (Mansournia et al., 2021)	La xarxa EQUATOR és una iniciativa internacional que pretén millorar la fiabilitat i el valor de la literatura d'investigació sanitària publicada promovent informes transparents i precisos i un ús més ampli de directrius d'informes amb rigor. Diferents guies segons els objectius marcats (CONSORT, STROBE, CHAMP, Ampl, TRIPOD,...) són necessaris per ajudar a pensar en molts aspectes del reporting i repassar aspectes latents d'inici en les anàlisis.
CONCLUSIONS			
Interpretació		(Spiegelhalter, 2019)	
Es connecten els resultats (gràfic, taula, resum dades) amb el coneixement existent del problema?	(Sí, No, No està clar, No procedeix)	(Arnold et al., 2021)	Cal pensar si la forma en què es presenten els resultats i la seva relació amb el que es sabia a priori té sentit.
Tenen sentit els resultats tenint en compte el coneixement actual?	(Sí, No, No està clar, No procedeix)	(Arnold et al., 2021) (Wild et al., 2018)	Cal pensar si els resultats obtinguts i la seva relació amb el que es sabia a priori té sentit.

S'han evitat anàlisis selectius o p-hacking?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	Cal conèixer o ser conscient del que és p-haching o Harking no només per fer un bon ús de les anàlisis sinó sobretot no caure amb males interpretacions.
S'han interpretat els resultats basant-se en mesures d'efecte?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	Cal interpretar els resultats basant-se en mesures d'associació i intervals de confiança o credibilitat (IC) i interpretar correctament els valors p grans com a resultats indecisos, per veure per exemple la no evidència d'absència de l'efecte.
S'ha distingit correctament la causalitat d'associació de la correlació?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	Correlació no implica causalitat no només en les anàlisis sinó sobretot en la interpretació de les troballes. Els resultats de les anàlisis preespecificades es distingeixen dels resultats de les anàlisis exploratòries en la interpretació.
S'ha sabut diferenciar la rellevància del camp amb la significació estadística?	(Sí, No, No està clar, No procedeix)	(Utts, 2021a)	Ens referim a "Significació pràctica vs estadística" que ens pot conduir al bon us i pensar què signifiquen per exemple els valors p o quan es poden treure conclusions causals.
Conclusions		(Spiegelhalter, 2019)	
S'és conscient de les limitacions a les conclusions basades en com es van mesurar i/o recopilar les dades?	(Sí, No, No està clar, No procedeix)	(GAISE College Report ASA Revision Committee, 2016)	Pensar en escriure o discutir les limitacions d'un estudi pot ajudar a prevenir "falses" conclusions (ex: tenir present validesa externa,...).

Les conclusions depenen de les fonts de les dades i seus anàlisis?	(Sí, No, No està clar, No procedeix)	(Utts, 2021a)	Fer-se aquestes preguntes poden dirigir a pensar en els punts forts i febles d'etapes anteriors del cicle i com depenen les conclusions d'aquestes.
Comunicació		(Spiegelhalter, 2019)	
És correcta la comunicació de la descripció de les dades?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	La mitjana i la desviació estàndard proporcionen un bon resum de les dades per variables contínues simètriques. L'error estàndard no és una bona opció per utilitzar-se en lloc de SD. En cas de dades quantitatives asimètriques, la mediana i el rang interquartil són més informatius. Les variables categòriques s'han de resumir en nombre i percentatge. Per dades longitudinals, s'hauria d'informar del temps de seguiment.
La comunicació en aquest estudi el fa reproduïble o replicable?	(Sí, No, No està clar, No procedeix)	(Schwabid et al., 2022; Serghiou et al., 2021)	Les preocupacions recents sobre la reproductibilitat de la ciència han donat lloc a diverses crides per a pràctiques de recerca més obertes. La transparència i l'obertura de la investigació són essencials per avaluar, replicar i implementar adequadament els resultats de la investigació
És acceptable la comunicació de la presentació de les figures i/o taules?	(Sí, No, No està clar, No procedeix)	(Mansournia et al., 2021)	Cal pensar si el format de taules i figures ajuda a comunicar els resultats obtinguts.
Es comunica la incertesa sobre fets, números i ciència?	(Sí, No, No està clar, No procedeix)	(van der Bles et al., 2019)	La incertesa és una part inherent del coneixement i moltes vegades s'evita comunicar obertament la seva incertesa sobre el que se sap, amb por de la reacció del seu públic. Tot i que hi ha algunes evidències que comunicar la incertesa epistèmica no necessàriament afecta

			negativament les audiències, l'impacte pot variar entre individus i formats de comunicació.
Es té clar a qui es comunica?	(Sí, No, No està clar, No procedeix)	(van der Bles et al., 2019)	Cal tenir present la comunicació i sobretot el seu públic per informar, motivar, instruir o influir en les persones. Els efectes de la comunicació d'incertesa depenen també de les característiques del públic objectiu i de la relació entre l'audiència i el comunicador, el tema o font de la incertesa. Les diferències importants entre els individus, com ara el seu nivell d'experiència, actituds prèvies, habilitats numèriques, nivell d'educació o optimisme, poden influir que la mateixa comunicació d'incertesa afecta les persones de manera diferent.
Impacte			
Hi ha un impacte polític?	(Sí, No, No està clar, No procedeix)	(Büttner et al., 2021)	L'impacte de les polítiques es refereix a la investigació que informa les regles establertes per una organització (és a dir, un responsable polític) per governar el comportament.
Hi ha un impacte econòmic?	(Sí, No, No està clar, No procedeix)	(Büttner et al., 2021)	Els impactes econòmics de la investigació en salut inclouen la comercialització de la investigació en salut aplicada, l'estalvi de costos sanitaris a través de la reducció de la morbiditat i la mortalitat com a resultat de les intervencions produïdes per la investigació en salut, o l'economia monetària que dona el valor de la millora de la salut que es basa en la investigació.

Hi ha un impacte social?	(Sí, No, No està clar, No procedeix)	(Büttner et al., 2021)	L'impacte social avarca molts termes com ara les activitats els beneficis socials, la utilitat social, el valor públic i la rellevància social de la investigació.
S'ha tingut en compte l'output acadèmic?	(Sí, No, No està clar, No procedeix)	(Büttner et al., 2021)	El rendiment acadèmic és una mesura del rendiment acadèmic i la productivitat de la recerca dels científics, i sovint es combina amb la importància i l'impacte de la investigació. El rendiment acadèmic es refereix a les contribucions intel·lectuals dels científics dins l'acadèmia. Moltes mètriques tenen com a objectiu capturar la producció acadèmica d'un científic i la seva investigació, incloent-hi el nivell de document (per exemple, el recompte de publicacions), el nivell d'autor (per exemple, el nombre d'afiliacions institucionals) i el nivell de revista (per exemple, factor d'impacte). No està clar, però, si la producció acadèmica es relaciona amb l'impacte de la recerca.

7. Referències annex

Arnold, P., & Franklin, C. (2021). What Makes a Good Statistical Question? Journal of Statistics and Data Science Education, 29(1). doi: 10.1080/26939169.2021.1877582

Booth, A., Noyes, J., Flemming, K., Moore, G., Tunçalp, Ö., & Shakibazadeh, E. (2019). Formulating questions to explore complex interventions within qualitative evidence synthesis. BMJ Global Health, 4. doi: 10.1136/bmjgh-2018-001107

Broman, K. W., & Woo, K. H. (2018). Data Organization in Spreadsheets. American Statistician, 72(1). doi: 10.1080/00031305.2017.1375989

Büttner, F., Ardern, C. L., Blazey, P., Dastouri, S., McKay, H. A., Moher, D., & Khan, K. M. (2021). Counting publications and citations is not just irrelevant: It is an incentive that subverts the impact of clinical research. In British Journal of Sports Medicine (Vol. 55, Issue 12). doi: 10.1136/bjsports-2020-103146

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. European Urology, 67(6). doi: 10.1016/j.eururo.2014.11.025

Cox, D. R., Kartsonaki, C., & Keogh, R. H. (2018). Big data: Some statistical issues. Statistics and Probability Letters, 136. doi: 10.1016/j.spl.2018.02.015

Davies, K. S. (2011). Evidence Based Library and Information Practice Commentary Formulating the Evidence Based Practice Question: A Review of the Frameworks. Evidence Based Library and Information Practice, 6(2).

GAISE College Report ASA Revision Committee. (2016). Guidelines for assessment and instruction in statistics education (GAISE) college report 2016. In Report.

Goldschmidt, G., & Matthews, B. (2022). Formulating design research questions: A framework. Design Studies, 78. doi: 10.1016/j.destud.2021.101062

Hidalgo, B., & Goodman, M. (2013). Multivariate or multivariable regression? In American Journal of Public Health (Vol. 103, Issue 1). doi: 10.2105/AJPH.2012.300897

Horton, N. J., & Hardin, J. S. (2021). Integrating Computing in the Statistics and Data Science Curriculum: Creative Structures, Novel Skills and Habits, and Ways to Teach Computational Thinking. Journal of Statistics and Data Science Education, 29(S1), S1–S3. doi: 10.1080/10691898.2020.1870416

Lang, T. A., & Altman, D. G. (2015). Basic statistical reporting for articles published in Biomedical Journals: The “Statistical analyses and methods in the published literature” or the SAMPL guidelines. In International Journal of Nursing Studies (Vol. 52, Issue 1). doi: 10.1016/j.ijnurstu.2014.09.006

Leek, J. T., & Peng, R. D. (2015). What is the question? Mistaking the type of question being considered is the most common error in data analysis. In Science (Vol. 347, Issue 6228). doi: 10.1126/science.aaa6146

MacKay, R. J., & Oldford, R. W. (2000). Scientific method, statistical method and the speed of light. Statistical Science, 15(3). doi: 10.1214/ss/1009212817

Mansournia, M. A., Collins, G. S., Nielsen, R. O., Nazemipour, M., Jewell, N. P., Altman, D. G., & Campbell, M. J. (2021). A CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): Explanation and elaboration. In British Journal of Sports Medicine (Vol. 55, Issue 18). doi: 10.1136/bjsports-2020-103652

McNamara, A., Horton, N. J., & Baumer, B. S. (2017). Greater Data Science at Baccalaureate Institutions. In Journal of Computational and Graphical Statistics (Vol. 26, Issue 4). doi: 10.1080/10618600.2017.1386568

Nielsen, R. O., Simonsen, N. S., Casals, M., Stamatakis, E., & Mansournia, M. A. (2020). Methods matter and the “too much, too soon” theory (part 2): What is the goal of your sports injury research? Are you describing, predicting or drawing a causal inference? In British Journal of Sports Medicine. doi: 10.1136/bjsports-2020-102144

Praharaj, S. K., & Ameen, S. (2020). How to choose research topic? Kerala Journal of Psychiatry, 33(1). doi: 10.30834/kjp.33.1.2020.188

Schwabid, S., Janiaud, P., Dayanid, M., Amrhein, V., Panczakid, R., Palagiid, P. M., Hemkensid, L. G., Ramonid, M., Rothenid, N., Sennid, S., Furrerid, E., & Heldid, L. (2022). Ten simple rules for good research practice. doi: 10.1371/journal.pcbi.1010139

Serghiou, S., Contopoulos-Ioannidis, D. G., Boyack, K. W., Riedel, N., Wallach, J. D., & Ioannidis, J. P. A. (2021). Assessment of transparency indicators across the

biomedical literature: How open is open? PLoS Biology, 19(3). doi: 10.1371/journal.pbio.3001107

Spiegelhalter, D. (2019). The art of statistics: Learning from data. .

Utts, J. (2021a). Enhancing Data Science Ethics Through Statistical Education and Practice. International Statistical Review, 89(1). doi: 10.1111/insr.12446

Utts, J. (2021b). Statistical Practice Is Not a Spectator Sport. Harvard Data Science Review. doi: 10.1162/99608f92.ff65fd7a

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). Communicating uncertainty about facts, numbers and science. Royal Society Open Science, 6(5). doi: 10.1098/rsos.181870

Vandenbroucke, J. P., & Pearce, N. (2018). From ideas to studies: How to get ideas and sharpen them into research questions. Clinical Epidemiology, 10. doi: 10.2147/CLEP.S142940

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. In PLoS Biology (Vol. 13, Issue 4). doi: 10.1371/journal.pbio.1002128

Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., Savic, M. D., Garovic, V. D., & Milic, N. M. (2019). Reveal, Don’t Conceal: Transforming Data Visualization to Improve Transparency. In Circulation (Vol. 140, Issue 18). doi: 10.1161/CIRCULATIONAHA.118.037777

Wickham, H., & Grolemund, G. (2017). R for Data Science: Visualize, Model, Transform, Tidy, and Import Data. In O’Reilly Media.

Wild, C. J., Utts, J. M., & Horton, N. J. (2018). What is statistics?. In International handbook of research in statistics education (pp. 5-36). Springer, Cham. (pp. 5–36).

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution, 1(1). doi: 10.1111/j.2041-210x.2009.00001.x