

TRABAJO FIN DE MÁSTER

MÁSTER UNIVERSITARIO OFICIAL EN CIENCIA DE DATOS E
INGENIERÍA DE COMPUTADORES

Epidemiología y detección de biomarcadores en cáncer de páncreas

Autor:

Daniel Redondo Sánchez

Tutores:

Ignacio Rojas

Luis Javier Herrera

Daniel Castillo

Granada, septiembre de 2020



**UNIVERSIDAD
DE GRANADA**

0. Índice general

1. Cáncer	5
1.1. ¿Qué es el cáncer?	5
1.2. Causas del cáncer	5
1.3. Impacto del cáncer	6
1.4. Cáncer de páncreas	7
2. Indicadores epidemiológicos del cáncer de páncreas.	9
2.1. Indicadores epidemiológicos	9
2.2. Incidencia	9
2.3. Mortalidad	16
2.4. Prevalencia	18
2.5. Supervivencia	18
3. Líneas abiertas de trabajo	19
Bibliografía	20
Anexo: Código en R	23

1. Cáncer

1.1. ¿Qué es el cáncer?

El cáncer es un conjunto de enfermedades que tienen en común una división incontrolada de las células, que se propagan a tejidos circundantes [1]. Casi cualquier parte del cuerpo humano, que está formado por unas $3 \cdot 10^{13}$ células [2], puede desarrollar cáncer. Por ello, aunque hablemos del cáncer como una única enfermedad, en realidad hay más de 100 tipos distintos de cáncer, con diferentes características.

1.2. Causas del cáncer

Las causas del cáncer son diversas, y el proceso de creación del cáncer es complejo y multifactorial. A menudo el causante del cáncer no es un solo elemento, sino la combinación e interacción de distintos factores ambientales y genéticos. Los factores ambientales son la causa del cáncer en un 90-95 % de los casos, mientras que la genética es responsable del restante 5-10 % [3], si bien estos porcentajes varían en función del tipo de cáncer.

Los factores causantes del cáncer se pueden clasificar principalmente en dos categorías: factores no modificables (esto es, factores sobre los que no tenemos ningún control) y factores modificables (donde se centra la prevención del cáncer).

Factores no modificables:

- Edad. La edad es el principal factor de riesgo del cáncer. En general, el cáncer aumenta con la edad, aunque hay algunos cánceres que son más frecuentes en niños o adolescentes.
- Sexo. El cáncer suele ser más frecuente en hombres que en mujeres.
- Historial familiar, en algunas localizaciones anatómicas como mama, colon o próstata.

- Enfermedades genéticas, como el síndrome de Lynch y su relación con el cáncer colorrectal.

Factores modificables:

- Tabaco. Aumenta el riesgo de incidencia de algún tipo de cáncer entre 10-20 veces más.
- Alcohol. El consumo de alcohol se asocia con una mayor incidencia de cáncer de localizaciones como estómago, mama, colon, o recto.
- Dieta, sedentarismo y peso corporal. Para prevenir el cáncer es importante llevar una dieta equilibrada, realizar ejercicio físico y evitar la obesidad.
- Exposición solar. Debe controlarse para disminuir el riesgo de melanoma y otros tumores de la piel.
- Exposición a distintos carcinógenos (agentes que producen cáncer). La *International Agency for Research on Cancer* (IARC) tiene registrados 120 agentes (químicos, metales, radiaciones, productos farmacéuticos ...) que producen cáncer en humanos [4], si bien diferentes exposiciones producen distintos tipos de cáncer [5].

Otros factores no se pueden clasificar como modificables o no modificables, ya que algunos de sus aspectos no se pueden cambiar. Los principales son:

- Factores socioeconómicos. Son aquellas características del entorno del sujeto, como nivel socioeconómico, grado de cobertura sanitaria y lugar de residencia. Las desigualdades socioeconómicas se reflejan a menudo en desigualdades en cáncer.
- Factores reproductivos y hormonales, como la toma de anticonceptivos orales, la lactancia, o la terapia hormonal sustitutiva en mujeres menopáusicas.

1.3. Impacto del cáncer

El cáncer es en la actualidad uno de los mayores problemas de Salud Pública a nivel mundial. En el año 2018 se diagnosticaron en el mundo 18,1 millones de casos nuevos de cáncer, y se produjeron 9,6 millones de defunciones por cáncer [6]. El cáncer de pulmón es el más diagnosticado (12 % del total), y la principal causa

de muerte por cáncer (18%). Existen importantes variaciones geográficas en la distribución del cáncer en el mundo, debidas principalmente a las diferencias en estilos de vida y al contexto económico y sanitario de las regiones [6].

En Europa, durante el año 2018 se han diagnosticado 3,9 millones de casos de cáncer (exceptuando el cáncer de piel no melanoma) y se han producido 1,9 millones de defunciones por cáncer [7].

Los últimos datos de incidencia disponibles para España son del año 2019, donde se estima que más de 277.000 personas sean diagnosticadas de cáncer [8], mientras que hubo más de 113.000 defunciones por cáncer en el año 2017 [9].

A la vista de estos datos, podemos afirmar que el estudio del cáncer en todas sus vertientes es necesario para reducir el impacto de la enfermedad en la población, mediante acciones como la planificación de la atención sanitaria, la evaluación de la efectividad de sus tratamientos, o el desarrollo de estrategias de prevención como los programas de cribado.

1.4. Cáncer de páncreas

2. Indicadores epidemiológicos del cáncer de páncreas.

2.1. Indicadores epidemiológicos

Para medir en la población el impacto del cáncer se utilizan principalmente cuatro factores:

- **Incidencia** (casos nuevos). Indica el riesgo de presentar cáncer.
- **Mortalidad** (defunciones). Indica el riesgo de morir por cáncer.
- **Prevalencia** (casos nuevos y antiguos, vivos). Indica la carga asistencial de la enfermedad.
- **Supervivencia** (porcentaje de casos vivos). Indica la historia natural del cáncer y la efectividad del tratamiento.

Estos factores son clave para la vigilancia epidemiológica, las actividades de prevención y la planificación de la asistencia sanitaria.

2.2. Incidencia

Número de casos

La medida básica para medir la incidencia de cáncer es el número de casos nuevos. A partir de este indicador, se calculan otros más complejos que proporcionan distinta información sobre la incidencia.

Tasa bruta.

La tasa bruta (TB) por 100.000 habitantes para un periodo concreto se define como el cociente entre el número de casos nuevos y el número de personas-año a riesgo por 100.000 [10]. El número de personas-año a riesgo es la población que podría haber sido diagnosticada de cáncer, y en general se calcula como la suma de la población año a año.

$$TB = \frac{\text{Número de casos nuevos}}{\text{Personas-año a riesgo}} \cdot 100.000 = \frac{N}{P} \cdot 100.000$$

Usualmente, este cociente se multiplica por un múltiplo de 10, para facilitar la interpretabilidad del dato. Si en el caso de la natalidad, la tasa bruta se suele informar por cada 1.000 habitantes, en el caso del cáncer lo usual es representar la tasa bruta por cada 100.000 habitantes.

Las tasas brutas se usan muy a menudo debido que son fácilmente calculables e interpretables, aunque pueden enmascarar diferencias existentes entre varios grupos de edad [11]. Para analizar la incidencia en distintas edades, se utilizan las tasas específicas por edad.

Tasa específica por edad

La tasa específica por edad (TEE) se define como la tasa bruta para un grupo de edad específico [10]. Esto es, la tasa específica por edad para el grupo de 30 a 34 años se calcula dividiendo el número de casos nuevos ocurridos en personas de 30 a 34 años entre el número de personas-año en ese mismo rango de edad.

Aunque las tasas específicas se pueden calcular para cualquier rango de edad, es común que se representen por grupos de edad quinquenales. En ocasiones, puede resultar de interés analizar por separado los indicadores para niños menores de 1 año, que pueden tener características diferentes al resto de niños, y agrupar a los mayores de una determinada edad en un sólo grupo de edad. Es por ello que, en general, las tasas específicas por edad se suelen calcular en 18 grupos de edad (0-4, 5-9, ..., 80-84, 85 y más), 19 grupos de edad (0, 1-4, 5-9, ..., 80-84, 85 y más) o 21 grupos de edad (0, 1-4, 5-9, ..., 80-84, 85-89, 90-94, 95 y más).

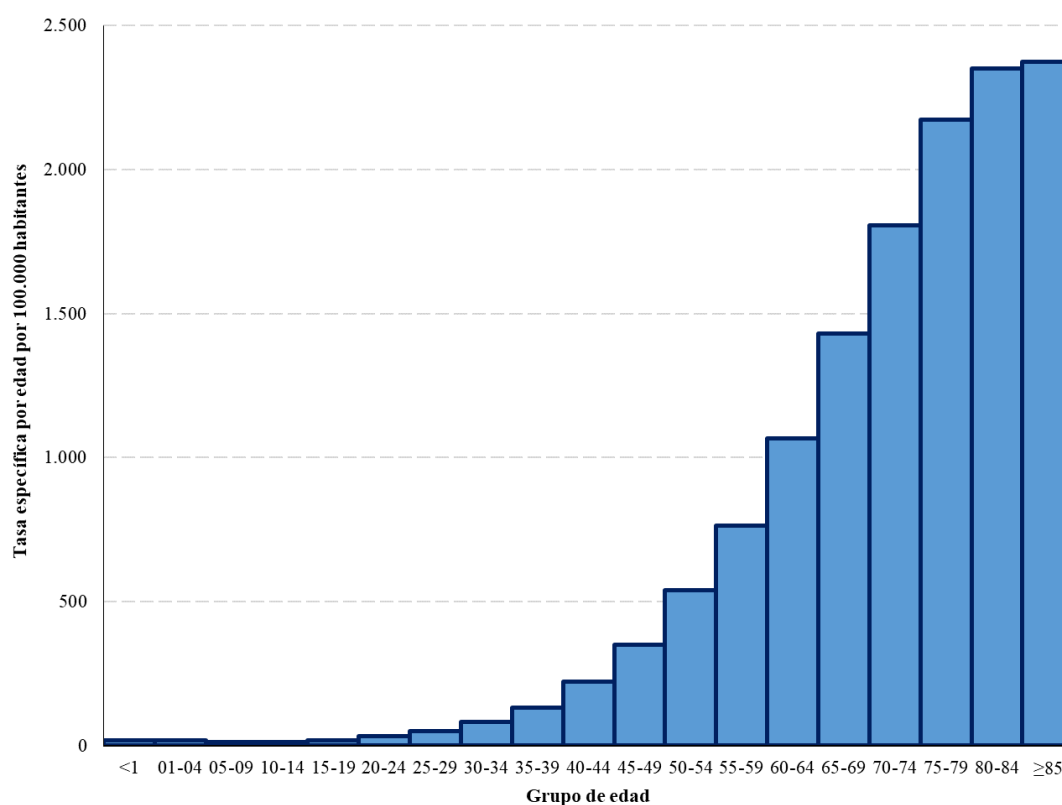
La tasa específica por edad para el i -ésimo grupo de edad viene dada por:

$$TEE_i = \frac{N_i}{P_i} \cdot 100.000$$

donde N_i indica el número de casos nuevos ocurridos en el i -ésimo grupo de edad, y P_i indica las personas-año en el i -ésimo grupo de edad.

En la Figura 3 se representan las tasas específicas por edad (19 grupos de edad) para el total del cáncer en la provincia de Granada para el periodo 1985-2013.

Figura 3. Tasas específicas por edad del total del cáncer (incluyendo piel no melanoma) en la provincia de Granada durante el periodo 1985-2013.



Tasa estandarizada por edad

Exceptuando algunos cánceres que son típicos en personas jóvenes, en general la frecuencia del cáncer aumenta con la edad. Este hecho hace que la comparación de tasas brutas entre diferentes poblaciones (o incluso entre la misma población en distintos periodos) no sea factible, dadas las diferencias en las estructuras de edad [12]. Esto es, si una población fuese mucho más joven que otra, incluso aunque las tasas específicas por edad sean iguales en ambas poblaciones, se diagnosticarían más casos en la población más anciana. Para resolver este inconveniente surgen las tasas estandarizadas por edad (ASR, por sus siglas en inglés:

Age-Standardised Rate).

Las tasas estandarizadas por edad son las tasas de incidencia que observaríamos en la población de estudio si esa población tuviese exactamente la misma estructura de edad que la población estándar predefinida [10]. Una población estándar consiste en la distribución de 100.000 habitantes ficticios en grupos de edad (generalmente, 18, 19 o 21 grupos de edad). Esta distribución teórica intenta reflejar la verdadera distribución de una determinada área geográfica, y permite la comparabilidad de tasas estandarizadas entre sus distintas regiones. El método más frecuente y recomendado de obtención de tasas estandarizadas por edad es el llamado “método directo” [13].

En la práctica, para el cálculo de una tasa estandarizada por edad se calculan primero las tasas específicas por edad. A continuación, la tasa de cada grupo de edad es multiplicada por un peso, que consiste en dividir entre 100.000 la población estándar de ese grupo de edad. Para N grupos de edad, la fórmula para obtener la tasa estandarizada por edad es:

$$ASR = \sum_{i=1}^N PEE_i \cdot TEE_i$$

donde TEE_i es la tasa específica por edad del i -ésimo grupo de edad y PEE_i es la población estándar para el i -ésimo grupo de edad.

Las poblaciones estándar más utilizadas son:

- Población mundial 1960. Propuesta por primera vez por Segi en 1960 [14], y modificada más tarde por Doll et al. en 1966 [15].
- Antigua población estándar europea (1976). Propuesta en 1976 por Waterhouse et al. [16] basándose en la estructura de edad de varias poblaciones escandinavas.
- Nueva población estándar europea (2013). Revisión de la población estándar europea de 1976 por parte de la Oficina Europea de Estadística (EUROSTAT), representa mejor el envejecimiento de la población europea [17]. El uso de esta población estándar aún no está ampliamente extendido, aunque se espera que en unos años sustituya completamente a su predecesora.

Además de redistribuir el peso de los determinados grupos de edad de la población estándar europea de 1976, permite la posibilidad de ampliar el número de grupos de edad a 21, añadiendo los grupos de edad 90-94 años y ≥ 95 años, adaptándose al aumento de habitantes muy ancianos.

Se notará ASR-W a la tasa estandarizada por la población mundial, y ASR-E a la tasa estandarizada por la población europea, especificando siempre el año de referencia (1976 o 2013).

En la Figura 4 se muestra una representación gráfica de las distintas poblaciones estándar descritas anteriormente, cuyos valores se encuentran en la Tabla 1.

Figura 4. Tasas específicas por edad del total del cáncer en la provincia de Granada durante el periodo 1985-2013.

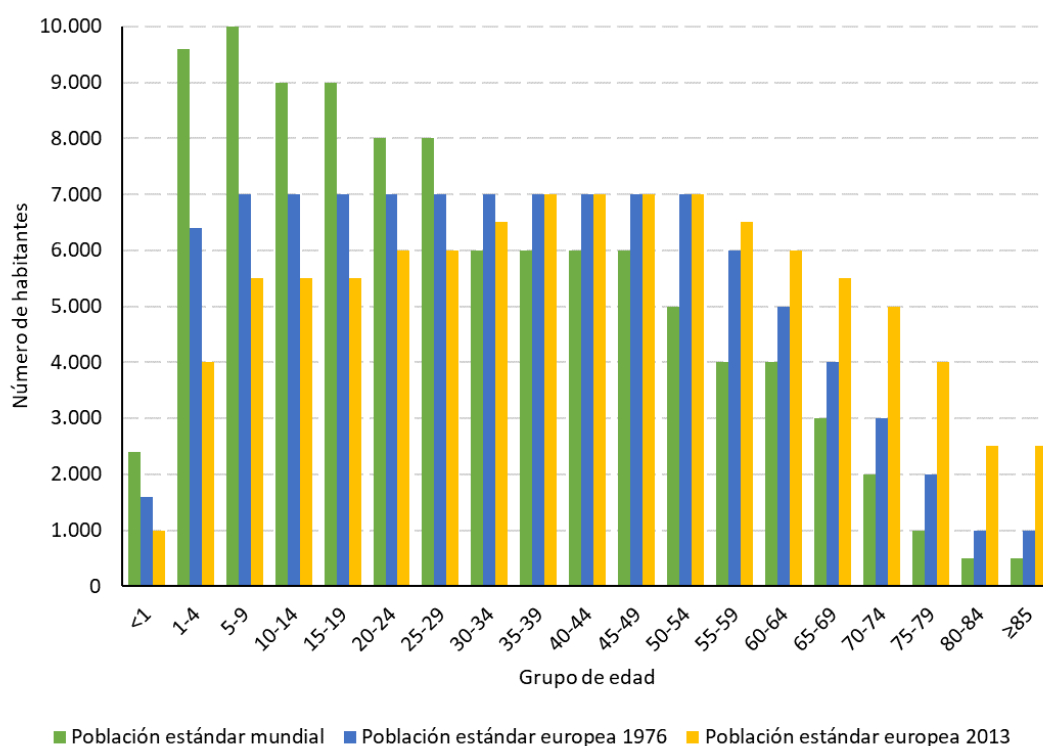


Tabla 1. Poblaciones estándar para el cálculo de tasas estandarizadas para 19 grupos de edad.

Grupo de edad	Población estándar mundial 1960	Población estándar europea 1976	Población estándar europea 2013
<1	2.400	1.600	1.000
1-4	9.600	6.400	4.000
5-9	10.000	7.000	5.500
10-14	9.000	7.000	5.500
15-19	9.000	7.000	5.500
20-24	8.000	7.000	6.000
25-29	8.000	7.000	6.000
30-34	6.000	7.000	6.500
35-39	6.000	7.000	7.000
40-44	6.000	7.000	7.000
45-49	6.000	7.000	7.000
50-54	5.000	7.000	7.000
55-59	4.000	6.000	6.500
60-64	4.000	5.000	6.000
65-69	3.000	4.000	5.500
70-74	2.000	3.000	5.000
75-79	1.000	2.000	4.000
80-84	500	1.000	2.500
≥85	500	1.000	2.500
Total	100.000	100.000	100.000

A continuación se muestra un ejemplo para el cálculo de tasas de incidencia: tasas brutas, específicas por edad y estandarizadas por edad.

Ejemplo 1. Incidencia de cáncer de próstata en hombres. Provincia de Granada, 2013.

Ejemplo de cálculo de tasas de incidencia: tasas específicas por edad (TEE), tasa bruta (TB) y tasa estandarizada por la población europea (ASR-E) tomando como referencia la población estándar europea de 1976 (PEE).

i	Grupo de edad	Población (P_i)	Número de casos (N_i)	TEE $_i$	PEE $_i$	TEE $_i$ * PEE $_i$
1	<1	4.500	0	0	1.600	0
2	1-4	20.075	0	0	6.400	0
3	5-9	27.120	0	0	7.000	0
4	10-14	25.726	0	0	7.000	0
5	15-19	25.237	0	0	7.000	0
6	20-24	28.262	0	0	7.000	0
7	25-29	30.438	0	0	7.000	0
8	30-34	36.044	0	0	7.000	0
9	35-39	38.508	0	0	7.000	0
10	40-44	36.546	0	0	7.000	0
11	45-49	36.850	8	21,7	7.000	1,5
12	50-54	32.594	17	52,2	7.000	3,7
13	55-59	26.740	35	130,9	6.000	7,9
14	60-64	21.895	67	306,0	5.000	15,3
15	65-69	19.737	88	445,9	4.000	17,8
16	70-74	15.123	92	608,3	3.000	18,3
17	75-79	14.092	75	532,2	2.000	10,6
18	80-84	10.671	47	440,4	1.000	4,4
19	≥ 85	6.807	18	264,4	1.000	2,6
						ASR-E = 82,1

Fuentes de información:

- Registro de Cáncer de Granada. Número de casos de cáncer de próstata en la provincia de Granada, 2013.
- Instituto Nacional de Estadística. Población de hombres residentes en la

provincia de Granada a 1 de Julio de 2013.

$$TEE_{14} = 100.000 * \frac{N_{14}}{P_{14}} = 100.000 * \frac{67}{21.895} = 306,0$$

La tasa específica por edad para los hombres de 60 a 64 años es de 306 casos por 100.000 hombres.

$$TB = 100.000 * \sum_{i=1}^{19} \frac{N_i}{P_i} = 100.000 * \frac{447}{456.965} = 97,8$$

La tasa bruta es de 97,8 casos por 100.000 hombres.

$$ASR-E = \sum_{i=1}^{19} TEE_i * PEE_i = 82,1$$

La tasa estandarizada por la población europea es de 82,1 casos por 100.000 hombres.

Otros indicadores

Existen otros indicadores epidemiológicos que miden distintos aspectos de la incidencia, como las tasas truncadas, las tasas acumulativas, o las tendencias de la incidencia. Más información sobre estos indicadores puede encontrarse en el capítulo 9 de “Registros de Cáncer: Principios y Métodos” (IARC) [10].

2.3. Mortalidad

Para la mortalidad, se pueden calcular de forma similar las mismas tasas que calculamos para la incidencia, cambiando el número de casos incidentes por el número de defunciones por cáncer.

Una defunción por cáncer se define como una muerte que es causada directamente por el cáncer, y no debe confundirse con defunciones de personas que tienen cáncer pero fallecen por otras causas (infarto, accidente de tráfico, ...).

A continuación se muestra un ejemplo para el cálculo de tasas de mortalidad: tasas brutas, específicas por edad y estandarizadas por edad.

Ejemplo 2. Mortalidad por cáncer de mama en mujeres. Provincia de Granada, 2013.

Ejemplo de cálculo de tasas de mortalidad: tasas específicas por edad (TEE), tasa bruta (TB) y tasa estandarizada por la población europea (ASR-E) tomando como referencia la población estándar europea de 1976 (PEE).

i	Grupo de edad	Población (P_i)	Número de defunciones (D_i)	TEE $_i$	PEE $_i$	TEE $_i$ * PEE $_i$
1	<1	4.149	0	0	1.600	0
2	1-4	18.380	0	0	6.400	0
3	5-9	25.345	0	0	7.000	0
4	10-14	23.548	0	0	7.000	0
5	15-19	23.674	0	0	7.000	0
6	20-24	26.741	0	0	7.000	0
7	25-29	29.494	0	0	7.000	0
8	30-34	34.431	0	0	7.000	0
9	35-39	36.199	4	11,1	7.000	0,8
10	40-44	35.204	5	14,2	7.000	1,0
11	45-49	36.540	6	16,4	7.000	1,1
12	50-54	32.853	11	33,5	7.000	2,3
13	55-59	27.334	19	69,5	6.000	4,2
14	60-64	23.124	9	38,9	5.000	1,9
15	65-69	21.835	8	36,6	4.000	1,5
16	70-74	17.918	12	67,0	3.000	2,0
17	75-79	18.563	12	64,6	2.000	1,3
18	80-84	15.643	14	89,5	1.000	0,9
19	≥ 85	12.954	23	177,6	1.000	1,8
						ASR-E = 18,8

Fuentes de información:

- Ministerio de Sanidad, Consumo y Bienestar Social. Número de defunciones por cáncer de mama en mujeres en la provincia de Granada, 2013.
- Instituto Nacional de Estadística. Población de mujeres residentes en la provincia de Granada a 1 de Julio de 2013.

$$TEE_{19} = 100.000 * \frac{N_{19}}{P_{19}} = 100.000 * \frac{23}{12.954} = 177,6$$

La tasa específica por edad de mortalidad para las mujeres de 85 años o más es de 177,6 defunciones por 100.000 mujeres.

$$TB = 100.000 * \sum_{i=1}^{19} \frac{N_i}{P_i} = 100.000 * \frac{123}{463.929} = 26,5$$

La tasa bruta de mortalidad es de 26,5 defunciones por 100.000 mujeres.

$$ASR-E = \sum_{i=1}^{19} TEE_i * PEE_i = 18,8$$

La tasa de mortalidad estandarizada por la población europea es de 18,8 defunciones por 100.000 mujeres.

2.4. Prevalencia

2.5. Supervivencia

3. Líneas abiertas de trabajo

Bibliografía

- [1] National Cancer Institute. What is Cancer?, 2015.
- [2] Ron Sender, Shai Fuchs, and Ron Milo. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*, 14(8):e1002533, aug 2016.
- [3] María José Sánchez Pérez and Juan Bayo Calero. Apuntes del curso Prevención y diagnóstico precoz del cáncer, 2016.
- [4] International Agency for Research on Cancer. Agents classified by the IARC Monographs. 1-123, 2019.
- [5] V. J. Coglianò, R. Baan, K. Straif, Y. Grosse, B. Lauby-Secretan, F. El Ghissassi, V. Bouvard, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet, and C. P. Wild. Preventable Exposures Associated With Human Cancers. *JNCI Journal of the National Cancer Institute*, 103(24):1827–1839, dec 2011.
- [6] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, nov 2018.
- [7] IARC, JRC, ENCR, and IACR. ECIS - European Cancer Information System, 2019.
- [8] Red Española de Registros de Cáncer (REDECAN). Estimaciones de la incidencia de cáncer en España, 2019. 2019.
- [9] Instituto Nacional de Estadística (INE). Estadísticas de defunciones según la causa de muerte, 2017.
- [10] IARC. *Registros de Cáncer: Principios y Métodos*. 1995.
- [11] World Health Organization) Silva, Isabel dos Santos (International Agency for Research on Cancer. *Cancer Epidemiology: Principles and Methods*. 1999.
- [12] Emanuele Crocetti, Tadek Dyba, Carmen Martos, Giorgia Randi, Roisin Rooney, and Manola Bettio. The need for a rapid and comprehensive adoption of the revised European standard population in cancer incidence comparisons. *European Journal of Cancer Prevention*, 26(5):447–452, sep 2017.

-
- [13] Hugh H. Wolfenden. On the Methods of Comparing the Moralities of Two or More Communities, and the Standardization of Death-Rates. *Journal of the Royal Statistical Society*, 86(3):399, may 1923.
 - [14] Segi M. Cancer mortality for selected sites in 24 countries (1950–57). *Sendai, Japan: Department of Public Health, Tohoku University of Medicine.*, 1960.
 - [15] Editors Doll R, Payne P, Waterhouse PAH. Cancer incidence in five continents, Volume I. *Geneva: Union Internationale Contre le Cancer.*, 1966.
 - [16] editors. Waterhouse JAH, Muir CS, Correa P, Powell J. Cancer incidence in five continents, Volume III. *Lyon: IARC*, page 3:456, 1976.
 - [17] EUROSTAT. Revision of the European standard population: Report of the Eurostat’s task force. Technical report, Luxembourg: European Union., 2013.

Anexo: Código en R

Anexo II: Código en R

Cabecera de fichero

Caja 1. Cabecera del fichero.

```
1 # 240 para 500 simulaciones, 24000 para 50000 simulaciones
2 ciclos.seleccionados <- 24000
3
4 ruta <- "RUTA"
5
6 año.a.estimar <- 2013
7
8 # Años para evolución RIM
9 total.años.evolucion.rim <- 15
10 añoFrim <- año.a.estimar - 5
11 añoIrim <- añoFrim - total.años.evolucion.rim + 1
12 # Por ejemplo:
13 # Si año.a.estimar es 2017, y se consideran 15 años de evolución
   que acaban en 2012, el año inicial será 1998
14
15 # Años NORDPRED (I = Año inicial, F = Año final, P = Periodo)
16 añoIP1 <- año.a.estimar - 22
17 añoFP1 <- año.a.estimar - 18
18 añoIP2 <- año.a.estimar - 17
19 añoFP2 <- año.a.estimar - 13
20 añoIP3 <- año.a.estimar - 12
21 añoFP3 <- año.a.estimar - 8
22 añoIP4 <- año.a.estimar - 7
23 añoFP4 <- año.a.estimar - 3
24
25 # P5 es el periodo del que obtenemos la mortalidad tras aplicar
   NORDPRED a los periodos del 1 al 4
26 añoIP5 <- año.a.estimar - 2
27 añoFP5 <- año.a.estimar + 2
```

Creación fichero de parametros.tumor

Caja 2. Definición de cuartiles en R.

```
1 cuartiles <- function(icdo)
```



```

2 {
3   # Se selecciona Granada (18 = 101) y los datos de los cuatro
      primeros periodos de NORDPRED (años añoIP1 hasta año añoFP4)
4   datos <- mortalidad[mortalidad$PROVINCIA==101 & mortalidad$ICDO
      ==icdo & mortalidad$AÑO>=añoIP1 & mortalidad$AÑO<=añoFP4,]
5   datos <- aggregate(datos[,4:22],list(ICDO=datos$ICDO),sum)
6   datos <- datos[,3:20]
7   suma <- sum(datos)
8   frecuencia <- (datos/suma)*100
9   delta1 <- 100
10  delta2 <- 100
11  delta3 <- 100
12  delta4 <- 100
13
14  # Se calculan P10, T1, T2 y P90.
15  # corteedad es el número del grupo de edad (de 1 a 18) que
      corresponde al P10.
16  for (i in 1:18) {
17    suma.sup <- sum(frecuencia[,1:i])
18    if (abs(10-suma.sup) < delta1) {
19      delta1 <- abs(10-suma.sup)
20      corteedad <- i
21      p10 <- i*5
22    }
23    if (abs(33.3333-suma.sup) < delta2) {
24      delta2 <- abs(33.3333-suma.sup)
25      t2 <- i*5
26    }
27    if (abs(66.6666-suma.sup) < delta3) {
28      delta3 <- abs(66.6666-suma.sup)
29      t3 <- i*5
30    }
31    if (abs(90-suma.sup) < delta4) {
32      delta4 <- abs(90-suma.sup)
33      p90 <- i*5
34    }
35  }
36  aux <- c(icdo, corteedad,p10,t2,t3,p90)
37  aux
38 }

```

Caja 3. Creación de `parametros.tumor` en R.

```

1 crear.parametros.tumor.RCG<-function(){
2   salida <- cuartiles(114)
3   salida <- rbind(salida, cuartiles(115))
4   salida <- rbind(salida, cuartiles(116))
5   salida <- rbind(salida, cuartiles(117))
6   salida <- rbind(salida, cuartiles(118))
7   salida <- rbind(salida, cuartiles(119))
8   salida <- rbind(salida, cuartiles(199))
9   salida <- rbind(salida, cuartiles(614))
10  salida <- rbind(salida, cuartiles(615))
11  salida <- rbind(salida, cuartiles(616))
12  salida <- rbind(salida, cuartiles(617))
13  salida <- rbind(salida, cuartiles(618))
14  salida <- rbind(salida, cuartiles(619))
15  salida <- rbind(salida, cuartiles(699))
16  # para respetar la estructura de parametros.tumor
17  subsalida1 <- salida[c(1:6), c(2:6)]
18  subsalida2 <- salida[7, c(2:6)]
19  subsalida3 <- salida[c(8:13), c(2:6)]
20  subsalida4 <- salida[14, c(2:6)]
21  cero5 <- rep(0, 5)
22  salida <- rbind(subsalida1, cero5, subsalida2, cero5,
23                  subsalida3, cero5, subsalida4, cero5)
24  icdo <- c(114, 115, 116, 117, 118, 119, 121, 199, 100, 614,
25            615, 616, 617, 618, 619, 621, 699, 600)
26  sexo <- c("M", "M", "M", "M", "M", "M", "M", "M", "M", "F", "F",
27            "F", "F", "F", "F", "F", "F", "F")
28  loc <- c("Colon", "Recto", "Pulmón", "Próstata", "Vejiga", "Est
29          ómago", "Colon-recto", "Otros", "Total excepto piel no
30          melanoma", "Colon", "Recto", "Pulmón", "Mama", "Cuerpo
31          uterino", "Ovario", "Colon-recto", "Otros", "Total excepto
          piel no melanoma")
32  tipo <- c("SP", "SP", "SP", "SP", "SP", "SP", "0", "SP", "0", "
33            SP", "SP", "SP", "SP", "SP", "SP", "0", "SP", "0")
34  escenario <- c("A", "A", "A", "A", "A", "A", "Especial", "A", "
35                Especial", "A", "A", "A", "A", "A", "A", "Especial", "A", "
36                Especial")
37  salida <- cbind(icdo, sexo, loc, salida, tipo, escenario)
38  cabecera <- c("CODIGO", "SEXO", "VALOR", "CORTEEDAD", "P10", "
39                T2", "T3", "P90", "TIPO", "ESCENARIO")
40  colnames(salida) <- cabecera
41  setwd(paste0(ruta, "\\datos"))

```

```

32 write.csv2(salida, "parametros.tumor.csv", row.names = FALSE,
           quote = FALSE)
33 print(paste("Se ha creado parametros.tumor.csv"), quote = FALSE
       )
34 }

```

Funciones para NORDPRED

Caja 4. Definición de las funciones `nordpred.estimate` y `nordpred.prediction`.

```

1 nordpred.estimate <- function(cases,pyr,noperiod,startestage,
  linkfunc="power5") {
2   Rplatform <- exists("R.Version")
3   if ( dim(cases)[1]!=18 || dim(pyr)[1]!=18 ) {
4     stop("\ncases\" and \"pyr\" must have data for 18 age groups"
5     )
6   }
7   if ( dim(cases)[2]>dim(pyr)[2]) {
8     stop("\npyr\" must include information about all periods in \"
9     cases\"")
10  }
11  if (dim(pyr)[2]==dim(cases)[2]) {
12    stop("\npyr\" must include information on future rates (E.g.
13    must be greater than \"cases\")")
14  }
15  if ((dim(pyr)[2]-dim(cases)[2])>5) {
16    stop("Package can not predict more than 5 periods (given by
17    sizes of \"pyr\" and \"cases\")")
18  }
19  if ((dim(cases)[2]-noperiod)<0) {
20    stop("More periods specified in \"noperiod\" than columns in
21    \"cases\"")
22  }
23  if (noperiod<3) {
24    stop("\nnoperiod\" must be 3 or larger to get enough for
25    estimating")
26  }
27  dnoperiods <- dim(cases)[2]
28  dnoagegr <- 18
29  ageno <- rep(1:dnoagegr,dnoperiods)
30  periodno <- sort(rep(1:dnoperiods,dnoagegr))
31  cohort <- max(ageno)-ageno+periodno

```

```

26 y <- c(as.matrix(pyr[,1:dnoperiods]))
27 apcdata <- data.frame(Cases= c(as.matrix(cases)),Age=ageno,
28 Cohort=cohort,Period=periodno,y=y)
29 apcdata <- apcdata[apcdata$Age>=startestage,]
30 apcdata <- apcdata[apcdata$Period>(dnoperiods-noperiod),]
31 if (Rplatform) {
32   y <- apcdata$y
33   power5link <- poisson()
34   power5link$link <- "0.2 root link Poisson family"
35   power5link$linkfun <- function(mu) { (mu/y)^0.2 }
36   power5link$linkinv <- function(eta) { pmax(.Machine$double.
37     eps, y*eta^5) }
38   power5link$mu.eta <- function(eta) { pmax(.Machine$double.
39     eps, 5*y*eta^4) }
40 } else {
41   y <- apcdata$y
42   power5link <- poisson()
43   power5link$link <- function(mu) { (mu/y)^0.2 }
44   power5link$inverse <- function(eta) { y*eta^(1/0.2) }
45   power5link$deriv <- function(mu) { 0.2*(1/y)*(mu/y)^(0.2 -
46     1) }
47 }
48 options(contrasts=c("contr.treatment","contr.poly"))
49 if (linkfunc=="power5") {
50   res.glm <- glm(Cases~as.factor(Age)+Period+as.factor(Period)+
51     as.factor(Cohort) -1,family=power5link,data=apcdata)
52 } else if (linkfunc=="poisson") {
53   res.glm <- glm(Cases~as.factor(Age)+Period+as.factor(Period)+
54     as.factor(Cohort)+ offset(log(y)) -1,family=poisson(),data
55     =apcdata)
56 } else {
57   stop("Unknown \"linkfunc\"")
58 }
59 if (Rplatform) {
60   pvalue <- 1-pchisq(res.glm$deviance,res.glm$df.residual)
61 } else {
62   pvalue <- 1-pchisq(res.glm$deviance,res.glm$df)
63 }
64 mod1 <- glm(Cases~as.factor(Age)+Period +as.factor(Cohort) +
65   offset(log(y)) -1,family=poisson,data=apcdata)
66 mod2 <- glm(Cases~as.factor(Age)+Period +I(Period^2)+as.factor(
67   Cohort) + offset(log(y)) -1,family=poisson,data=apcdata)
68 if (Rplatform) {

```

```

60   pdiff<-anova(mod1,mod2,test="Chisq")$"P(>|Chi|)"[2]
61   # Correction added 2015-09-19 for compabilty with newer R
      versions:
62   if (is.null(pdiff))
63     pdiff <- anova(mod1,mod2,test="Chisq")$"Pr(>Chi)"[2]
64 } else {
65   pdiff<-anova(mod1,mod2,test="Chisq")[2,7]
66 }
67 if (pdiff <0.05) {
68   suggestionrecent <-T
69 } else {
70   suggestionrecent <-F
71 }
72 res <- list(glm=res.glm,cases=cases,pyr=pyr,noperiod=noperiod,
      gofpvalue=pvalue,startestage=startestage,
73             suggestionrecent=suggestionrecent,pvaluerecent=
      pdiff,linkfunc=linkfunc)
74 class(res) <- "nordpred.estimate"
75 attr(res,"Call") <- sys.call()
76 return(res)
77 }
78
79 nordpred.prediction <- function(nordpred.estimate.object,
      startuseage,recent,cuttrend=c(0,.25,.5,.75,.75)) {
80   if (class(nordpred.estimate.object)!="nordpred.estimate") {
81     stop("Variable \"nordpred.estimate.object\" must be of type \"
      \"nordpred.estimate\"")
82   }
83   if (nordpred.estimate.object$startestage>startuseage) {
84     stop("\"startuseage\" is set to high compared to \"
      startestage\" in \"nordpred.estimate.object\"")
85   }
86   if (length(cuttrend)<(dim(nordpred.estimate.object$pyr)[2]-dim(
      nordpred.estimate.object$cases)[2])) {
87     err <- paste("\"cuttrend\" must always be at least the same
      length as")
88     err <- paste(err,"the number of periods with population
      forecasts")
89     stop(err)
90   }
91   cases <- nordpred.estimate.object$cases
92   pyr <- nordpred.estimate.object$pyr
93   noperiod <- nordpred.estimate.object$noperiod

```

```

94  nototper <- dim(pyr)[2]
95  noobsper <- dim(cases)[2]
96  nonewpred <- nototper-noobsper
97  if (length(cuttrend)>nonewpred) {
98    cuttrend <- cuttrend[1:nonewpred]
99  }
100 if (is.data.frame(pyr)) {
101   years <- names(pyr)
102 } else {
103   if (is.null(dimnames(pyr))) {
104     years <- paste("Periode",1:nototper)
105   } else {
106     years <- dimnames(pyr)[[2]]
107   }
108 }
109 datatable <- matrix(NA,18,nototper)
110 datatable[,1:(nototper-nonewpred)] <- as.matrix(cases)
111 datatable <- data.frame(datatable)
112 row.names(datatable) <- c("0-4","5-9","10-14","15-19","20-24","
    25-29","30-34","35-39","40-44","45-49","50-54","55-59","
    60-64","65-69","70-74","75-79","80-84","85+")
113 names(datatable) <- years
114 for (age in 1:(startuseage-1)) {
115   obsinc <- cases[age,(noobsper-1):noobsper]/pyr[age,(noobsper
    -1):noobsper]
116   if (sum(is.na(obsinc))) {
117     obsinc[is.na(obsinc)] <- 0
118   }
119   datatable[age,(noobsper+1):nototper] <- ((obsinc[,1]+obsinc
    [,2])/2)*pyr[age,(noobsper+1):nototper]
120 }
121 for (age in startuseage:18) {
122   startestage <- nordpred.estimate.object$startestage
123   coefficients <- nordpred.estimate.object$glm$coefficients
124   coh <- (18-startestage) - (age-startestage) + (noperiod
    +1:nonewpred)
125   noages <- 18-startestage+1
126   driftmp <- cumsum(1-cuttrend)
127   cohfind <- noages + (noperiod-1) + 1 + (coh-1)
128   maxcoh <- 18 - startuseage + noperiod
129   agepar <- as.numeric(coefficients[age-startestage+1])
130   driftfind <- pmatch("Period",attributes(coefficients)$names)
131   driftpar <- as.numeric(coefficients[driftfind])

```

```

132   cohpar <- rep(NA,length(coh))
133   for (i in 1:length(coh)) {
134     if (coh[i] < maxcoh) {
135       cohpar[i] <- as.numeric(coefficients[cohfind[i]])
136     } else {
137       # For young cohorts not estimated:
138       cohpar[i] <- as.numeric(coefficients[length(coefficients)
139         -(startuseage-startestage)])
140       cohpar[i][is.na(cohpar[i])] <- 0
141     }
142   }
143   if (recent) {
144     lpfind <- driftfind + noperiod-2
145     lppar <- as.numeric(coefficients[lpfind])
146     driftrecent <- driftpar - lppar
147   }
148   if (nordpred.estimate.object$linkfunc=="power5") {
149     if (recent) {
150       rate <- (agepar+driftpar*noobsper+driftrecent*driftmp+
151         cohpar)^5
152     } else {
153       rate <- (agepar+driftpar*(noobsper+driftmp)+cohpar)^5
154     }
155   } else { # Possion link:
156     if (recent) {
157       rate <- exp(agepar+driftpar*noobsper+driftrecent*driftmp+
158         cohpar)
159     } else {
160       rate <- exp(agepar+driftpar*(noobsper+driftmp)+cohpar)
161     }
162   }
163   datatable[age,(noobsper+1):nototper] <- rate*pyr[age,(
164     noobsper+1):nototper]
165 }
166 res <- list(predictions=datatable,pyr=pyr,nopred=nonewpred,
167   noperiod=nordpred.estimate.object$noperiod,
168   gofpvalue=nordpred.estimate.object$gofpvalue,recent
169     =recent,pvaluerecent=nordpred.estimate.object$
170     pvaluerecent,
171   cuttrend=cuttrend,startuseage=startuseage,
172     startestage=nordpred.estimate.object$startestage
173   ,
174   glm=nordpred.estimate.object$glm)

```

```

166 class(res) <- "nordpred"
167 attr(res,"Call") <- sys.call()
168 return(res)
169 }

```

Funciones para la validación de estimaciones

Caja 5. Definición de `funcion.ajuste`, que mide el MAPE en R.

```

1 funcion.ajuste<-function(localizacion, escenario, sims = 100,
2   nivelconfianza, imprimir = "FALSE"){
3   if(escenario=="TODOS"){
4     salida <- funcion.ajuste(localizacion,"E3",sims,
5       nivelconfianza)
6     salida <- rbind(salida, funcion.ajuste(localizacion, "E1",
7       sims, nivelconfianza))
8     salida <- rbind(salida, funcion.ajuste(localizacion, "E5",
9       sims, nivelconfianza))
10    salida <- rbind(salida, funcion.ajuste(localizacion, "L",
11      sims, nivelconfianza))
12    salida <- rbind(salida, funcion.ajuste(localizacion, "C",
13      sims, nivelconfianza))
14    salida
15  }
16  else{
17    # 121 = Colon-recto en hombres
18    # 621 = Colon-recto en mujeres
19    if(localizacion != 121 && localizacion != 621){
20      observado <- estimado.observado(localizacion, escenario,
21        sims, nivelconfianza)$observado
22      estimado <- estimado.observado(localizacion, escenario,
23        sims, nivelconfianza)$estimado
24    }
25    else{
26      if(localizacion==121){
27        observado.colon <- estimado.observado(114, escenario,
28          sims, nivelconfianza)$observado
29        observado.recto <- estimado.observado(115, escenario,
30          sims, nivelconfianza)$observado

```



```

24     observado <- observado.colon+observado.recto
25     estimado.colon <- estimado.observado(114, escenario, sims
26         , nivelconfianza)$estimado
27     estimado.recto <- estimado.observado(115, escenario, sims
28         , nivelconfianza)$estimado
29     estimado <- estimado.colon+estimado.recto
30 }
31 else{
32     observado.colon <- estimado.observado(614, escenario,
33         sims, nivelconfianza)$observado
34     observado.recto <- estimado.observado(615, escenario,
35         sims, nivelconfianza)$observado
36     observado <- observado.colon+observado.recto
37     estimado.colon <- estimado.observado(614, escenario, sims
38         , nivelconfianza)$estimado
39     estimado.recto <- estimado.observado(615, escenario, sims
40         , nivelconfianza)$estimado
41     estimado <- estimado.colon+estimado.recto
42 }
43 }
44
45 if(imprimir == "TRUE"){
46     print("Observado:", quote = "FALSE")
47     print(observado)
48     print("Estimado:", quote = "FALSE")
49     print(estimado)
50 }
51
52 # MAPE
53 f <- 100 * (1 / 10) * sum(abs((estimado - observado) /
54     observado))
55 icdo <- parametros.tumor[parametros.tumor$CODIGO ==
56     localizacion, "VALOR"]
57 data.frame(localizacion, icdo, escenario, f)
58 }
59 }
60 }

```

Caja 6. Definición de las funciones `mejor.escenario.ajuste`, `funcion.ajuste.total` y `tabla.mejores.escenarios`.

```

1 mejor.escenario.ajuste <- function(localizacion, sims = 100,
  nivelconfianza, imprimir = "FALSE"){
2   esc.e1 <- funcion.ajuste(localizacion, "E1", sims,
    nivelconfianza, "FALSE")[[4]]
3   esc.e3 <- funcion.ajuste(localizacion, "E3", sims,
    nivelconfianza, "FALSE")[[4]]
4   esc.e5 <- funcion.ajuste(localizacion, "E5", sims,
    nivelconfianza, "FALSE")[[4]]
5   esc.l <- funcion.ajuste(localizacion, "L" , sims,
    nivelconfianza, "FALSE")[[4]]
6   esc.c <- funcion.ajuste(localizacion, "C" , sims,
    nivelconfianza, "FALSE")[[4]]
7
8   #Vemos cuál es el mejor escenario
9   if(esc.e1<esc.e3 && esc.e1<esc.e5 && esc.e1<esc.l && esc.e1<esc
    .c) esc.mejor="E1" else{
10    if(esc.e3<esc.e1 && esc.e3<esc.e5 && esc.e3<esc.l && esc.e3<
        esc.c) esc.mejor="E3" else{
11    if(esc.e5<esc.e1 && esc.e5<esc.e3 && esc.e5<esc.l && esc.e5
        <esc.c) esc.mejor="E5" else{
12    if(esc.l<esc.e1 && esc.l<esc.e5 && esc.l<esc.e3 && esc.l<
        esc.c) esc.mejor="L" else{
13    esc.mejor="C"
14    }
15    }
16    }
17  }
18
19  ifelse(imprimir == "FALSE",
20    salida <- paste0(funcion.ajuste(localizacion, esc.mejor,
        sims, nivelconfianza, "FALSE")[[3]]),
21    salida <- funcion.ajuste(localizacion, esc.mejor, sims,
        nivelconfianza, "FALSE"))
22  salida
23 }
24
25 funcion.ajuste.total <- function(sims = 100, nivelconfianza){
26
27   # Hombres
28   salida <- funcion.ajuste(114,"TODOS",sims,nivelconfianza)
29   salida <- rbind(salida, funcion.ajuste(115, "TODOS", sims,
        nivelconfianza))
30   salida <- rbind(salida, funcion.ajuste(116, "TODOS", sims,

```

```

    nivelconfianza))
31 salida <- rbind(salida, funcion.ajuste(117, "TODOS", sims,
    nivelconfianza))
32 salida <- rbind(salida, funcion.ajuste(118, "TODOS", sims,
    nivelconfianza))
33 salida <- rbind(salida, funcion.ajuste(119, "TODOS", sims,
    nivelconfianza))
34 salida <- rbind(salida, funcion.ajuste(199, "TODOS", sims,
    nivelconfianza))
35
36 # Mujeres
37 salida <- rbind(salida, funcion.ajuste(614, "TODOS", sims,
    nivelconfianza))
38 salida <- rbind(salida, funcion.ajuste(615, "TODOS", sims,
    nivelconfianza))
39 salida <- rbind(salida, funcion.ajuste(616, "TODOS", sims,
    nivelconfianza))
40 salida <- rbind(salida, funcion.ajuste(617, "TODOS", sims,
    nivelconfianza))
41 salida <- rbind(salida, funcion.ajuste(618, "TODOS", sims,
    nivelconfianza))
42 salida <- rbind(salida, funcion.ajuste(619, "TODOS", sims,
    nivelconfianza))
43 salida <- rbind(salida, funcion.ajuste(699, "TODOS", sims,
    nivelconfianza))
44
45 salida
46
47 }
48
49 tabla.mejores.escenarios<-function(sims=100,nivelconfianza){
50   v<-funcion.ajuste.total(sims,nivelconfianza)
51   v<-v[order(v$f),]
52
53   # Hombres
54   salida <- t(subset(v, v$localizacion == "114"))
55   salida <- rbind(salida, t(subset(v, v$localizacion == "115")))
56   salida <- rbind(salida, t(subset(v, v$localizacion == "116")))
57   salida <- rbind(salida, t(subset(v, v$localizacion == "117")))
58   salida <- rbind(salida, t(subset(v, v$localizacion == "118")))
59   salida <- rbind(salida, t(subset(v, v$localizacion == "119")))
60   salida <- rbind(salida, t(subset(v, v$localizacion == "199")))
61

```

```
62 # Mujeres
63 salida <- rbind(salida, t(subset(v, v$localizacion == "614")))
64 salida <- rbind(salida, t(subset(v, v$localizacion == "615")))
65 salida <- rbind(salida, t(subset(v, v$localizacion == "616")))
66 salida <- rbind(salida, t(subset(v, v$localizacion == "617")))
67 salida <- rbind(salida, t(subset(v, v$localizacion == "618")))
68 salida <- rbind(salida, t(subset(v, v$localizacion == "619")))
69 salida <- rbind(salida, t(subset(v, v$localizacion == "699")))
70
71 print(salida, quote = "FALSE")
72 }
```