Ambroa Rey, Eva

No publicación de resultados de los ensayos clínicos

1. Conseguir datos.

Para conseguir los datos accedemos a la web http://www.clinicaltrials.gov/. Aproximadamente a día de hoy hay registrados unos 64.000 estudios en oncología, de los cuáles aproximadamente 3.000 son españoles.
Esta web permite obtener los registros de los ensayos clínicos de manera muy sencilla. Extraemos los datos en un archivo csv con todas la columnas disponibles para su posterior procesado y análisis con la herramienta estadística R.
Debido a que el Acta del 2007 de la FDA requiere que las partes responsables de todos los ensayos de clínicos presenten los resultados resumidos en la base de datos de www.clinicaltrials.gov 12 meses después de su fecha de finalización primaria (PCD), seleccionaremos solamente los estudios completados, con una fecha de inicio (Study Start) posterior al 1 de Enero de 2008 y con una PCD (primary completion date) anterior al 31/10/2017. Así podremos comprobar si estos estudios han presentados resultados no más tarde del 31/10/2018.
Además, solamente seleccionaremos los estudios de tipo 'Interventional' y con fase 'Phase1 Phase2, Phase3 o Phase4'.
A continuación se muestra una imagen de la búsqueda de estos estudios.

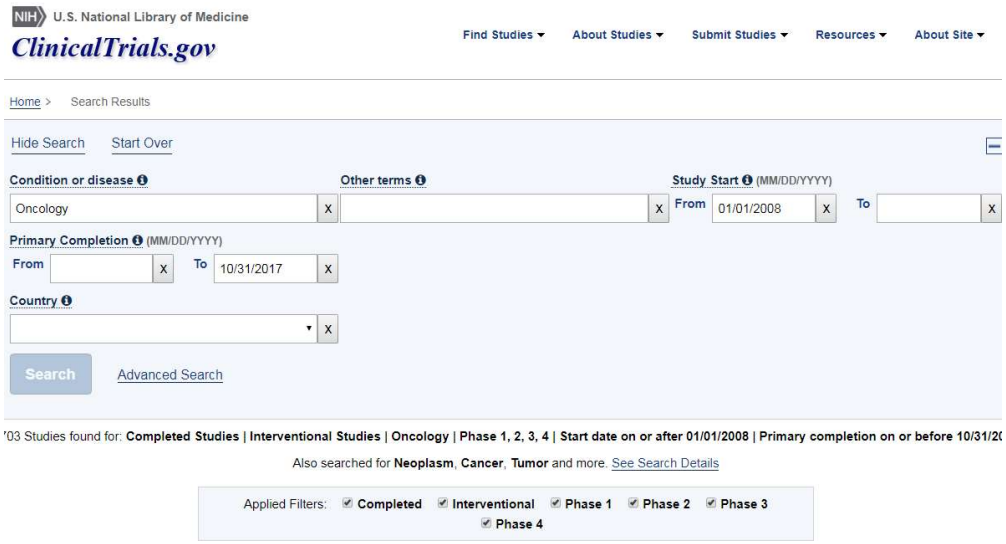

Imagen I. Búsqueda de los datos.

Descargamos esto datos en un csv y seleccionamos las columnas que nos interesan, que en este caso son las siguientes:

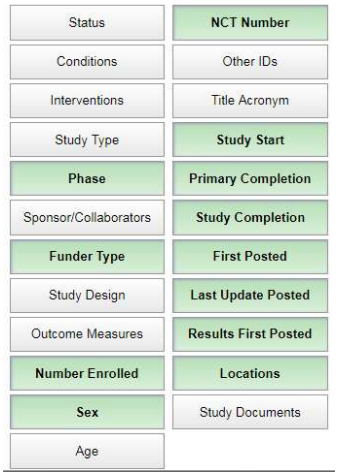

Imagen II. Selección de la información.

Con esto nos quedaremos con 7703 estudios (Datos.csv).
A parte del fichero que contiene todos los datos, también extraemos un fichero con los datos relevantes solamente de los estudios españoles, para hacer un análisis más profundo de estos datos. Este fichero contiene 696 estudios (Datos_Spain.csv).

2. Aplicar mecanismos de selección y filtrado de los datos.

Una vez descargados los datos abrimos el fichero 'Datos' haciendo uso del entorno de programación libre R.
Lo primero que haremos será hacer una evaluación rápida de los datos y eliminar las columnas que no necesitamos, que en este caso son 'Title', 'NCT number', 'Enrollment' y 'URL'.
El primer problema con el que nos encontramos son los valores nulos, hacemos una evaluación y encontramos 4 valores nulos en el campo 'Gender' y 322 en el campo 'Locations'. Ya que no podemos reemplazar estos valores nulos con otro valor (porque no tiene sentido rellenar con valor más probable en este caso), lo más adecuado será eliminar estos datos del análisis.

El segundo problema es que los datos no son consistentes en el formato. Por ejemplo, los campos que contienen fecha, no siguen un único formato, lo que dificulta la selección y limpieza de los datos. Para ello unificamos el formato de fecha, y creamos una nueva columna llamada 'DaysInterval' que corresponde a la diferencia en días entre la Primary Completion Date y la fecha en que se publicaron los primeros resultados.
El campo 'Funder Type' no es homogéneo, los nombres no aparecen en el mismo orden lo que nos da un resultado erróneo al evaluar los niveles que lo forman. Lo que haremos será unificar esos nombres, por ejemplo, Other|Industry lo cambiaremos a Industry|Other para que coincidan los formatos. También eliminaremos los niveles que contengan menos de 10 estudios para facilitar el análisis.
En el campo 'Locations', los diferentes centros que forman parte del estudio no están separados correctamente y es complicado evaluar cuántos centros forman parte del estudio.
El país no viene detallado en ninguna columna, forma parte de la columna 'Locations' por lo que tendremos que hacer un bucle que recorra este campo y extraiga el país de cada centro que participa en el estudio. Para ello, crearemos el

campo 'Country', donde aparecerá el nombre del país que registra el ensayo clínico y si son varios simplemente aparecerá 'Multicentric' ya que en algunos casos existen más de 20 países involucrados y es complicado analizar estos datos.
A continuación creamos otra columna llamada 'Multicentric' que tendrá dos valores, si el estudio es multicéntrico asignaremos el valor 1 y si no lo es asignaremos el valor 0. Posteriormente también eliminaremos la columna 'Locations'.
Después de filtrar y limpiar los datos, nos quedaremos con 7369 observaciones y 12 variables.
Los datos limpios y pre-procesados se guardarán en un fichero llamado 'Datos_finales'. El repositorio de github donde se almacenarán los datos y el código R ('Limpieza_Datos.R') y un PDF con los pasos seguidos es el siguiente:
https://github.com/eambroa/Ensayos-clinicos

Para los datos de estudios españoles, 'Datos_Spain.csv' se hará algo similar ya que los datos presentan los mismos problemas que los datos anteriores. Se añadirá para estos datos una variable nueva que recoja la ciudad ('City') donde se ha realizado el estudio. El código R, se llama 'Limpieza_Datos_Spain.R' y genera un fichero llamado 'Datos_Spain_final.csv'.

3. Extraer características.

Después de seleccionar y filtrar los datos, podemos extraer las primeras características. En la siguiente tabla resumimos algunas de ellas:

Característica	**Clasificación**	**Porcentaje**
Study Results	With results 2262 Without results: 5107	30.7% 69.3%
Gender	Both: 6004 Male: 441 Female: 924	81.5% 6.0% 12.5%
Study Phase	Phase1: 2384 Phase1/Phase2: 710 Phase2: 2884 Phase2/Phase3: 115 Phase3: 906 Phase 4: 370	32.3% 9.6% 39.1% 1.6% 12.3% 5.0%
Funded by	Industry: 3186 Industry/NIH: 15 Industry/Other: 1082 Industry/Other/NIH: 53 NIH: 331 NIH/Other:424	 43.2% 0.2% 14.7% 0.7%

	Other: 2278	4.6% 5.7% 30.9%

Tabla I. Características de los datos.

De la tabla anterior, lo más destacable es que un gran porcentaje de estudios no registra sus resultados en la web (aproximadamente un 70%). Podemos intentar evaluar si esto depende del tipo de fundación, de si es un estudio multicéntrico o no, de la fase del estudio, del género, etc.

Podemos graficar algunos resultados para ver si hay alguna tendencia clara:

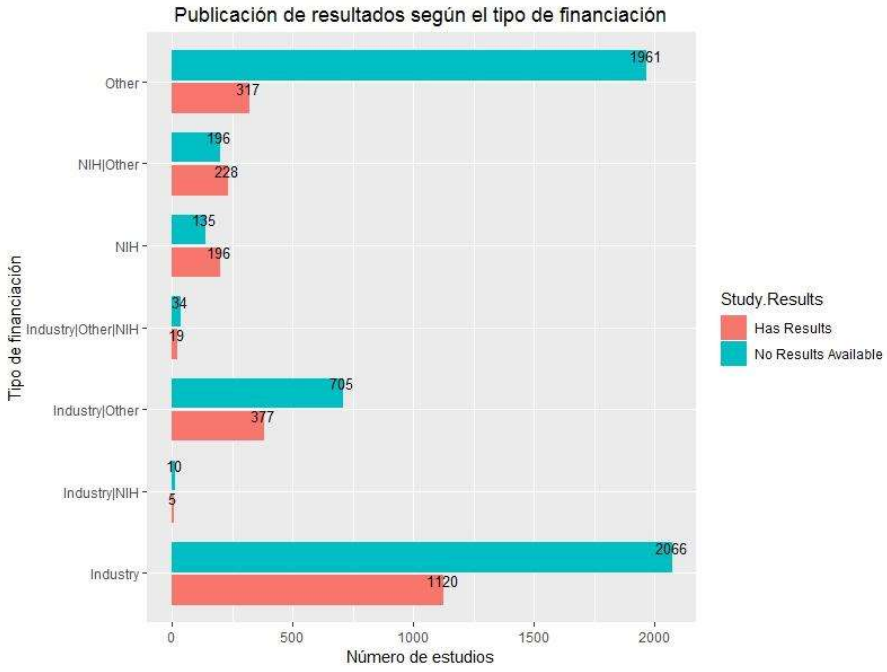

Imagen III. Resultados según la financiación

De la Imagen III, se puede ver, por ejemplo, que los estudios con financiación tipo 'Other' son los que menos resultados publican en la web.

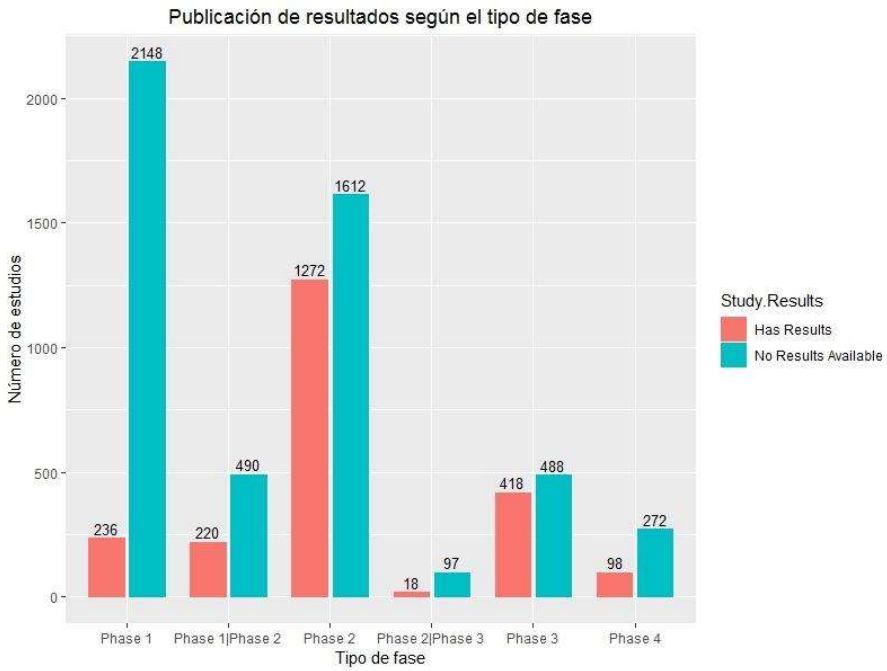

Imagen IV. Resultados según la fase

En la Imagen IV se observa que los estudios Fase1 son lo que menos resultados publican, seguido de los estudios Fase4.

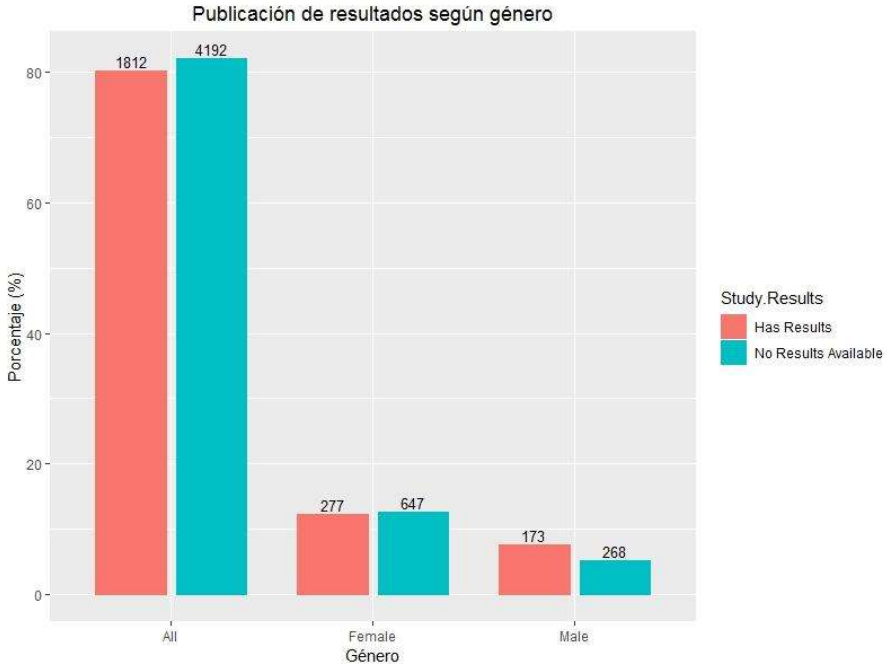

Imagen V. Resultados según el género

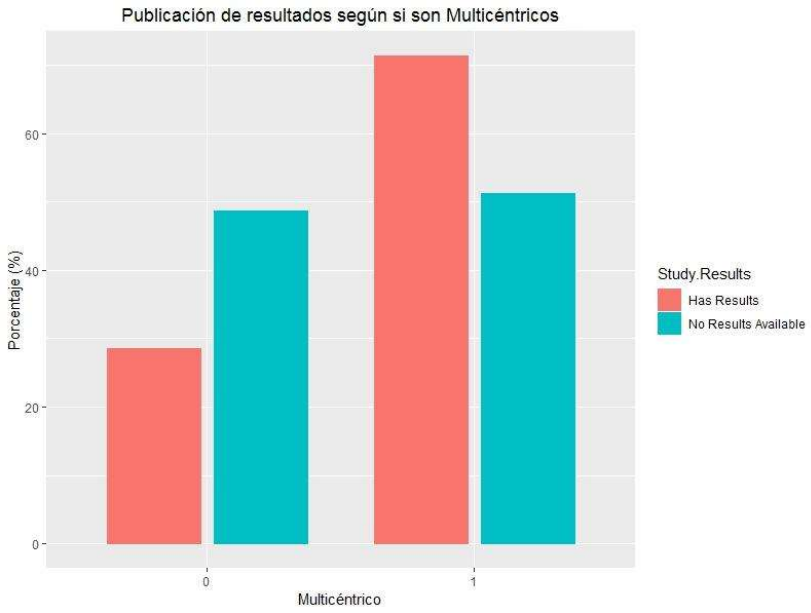

Imagen VI. Resultados según el centro

En la Imagen VI, se puede ver que los estudios multicéntricos publican más resultados que los estudios que se llevan a cabo en un solo centro.
A continuación analizaremos las características de los ensayos clínicos atendiendo a la publicación de resultados:

- *Qué tienen en común los ensayos clínicos que publican resultados?*

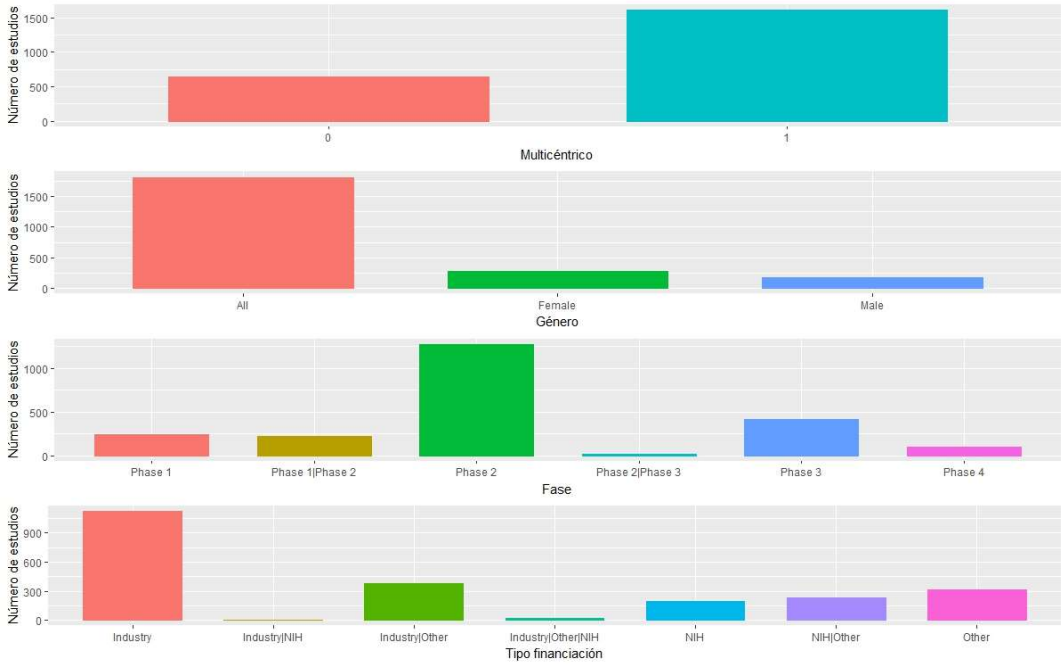

Imagen VII. Características de los ensayos clínicos que publican resultados

De la imagen anterior, podemos destacar que los estudios con resultados publicados son mayoritariamente financiados por la industria, incluyen ambos géneros, los más realizados son los estudios Fase2 y multicéntricos.

- *Qué características tienen los ensayos clínicos que no publican resultados?*

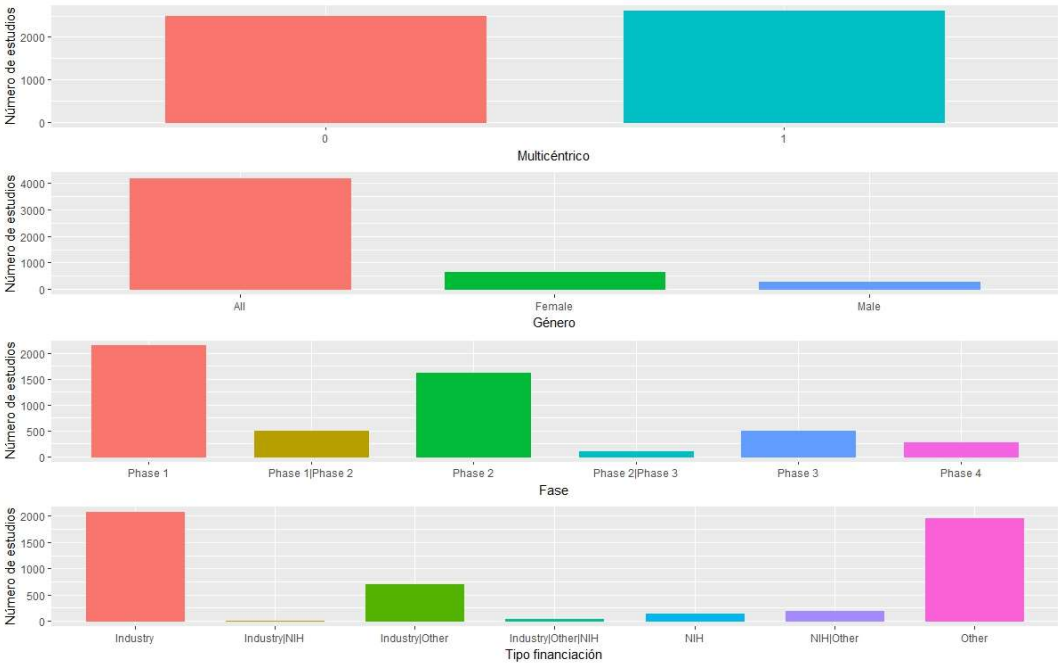

Imagen VIII. Características de los ensayos clínicos que no presentan resultados

Las principales características de los ensayos clínicos que no publican resultados son que están financiados por la industria u otros (universidades), que la gran mayoría son de tipo Fase1 y Fase2, incluyen a ambos géneros y no se diferencian en cuanto a la característica 'Multicéntrico'.

Tenemos que analizar también si los resultados publicados se hicieron dentro de los 12 meses posteriores a la PCD (primary completion date). Para esto evaluamos solamente los estudios que tienen resultados publicados y comprobamos que la fecha en la que se publicaron entra dentro del margen de los 12 meses.

De los 2262 estudios con resultados encontramos lo siguiente:

- PCD-fecha resultados <366 días: 11% (248 estudios)
- Media de días desde la PCD hasta la publicación de resultados: 217.7 días

4. Descriptores estadísticos

Observando las imágenes anteriores, parece existir una relación entre varias variables. Como tenemos variables categóricas, para evaluar si estas diferencias son estadísticamente significativas utilizaremos la prueba de chi-cuadrado con un 95% de intervalo de confianza. Las hipótesis a contrastar son las siguientes:
Hipótesis nula: la publicación resultados no depende de ningún parámetro.
Hipótesis alternativa 1: la publicación de resultados depende del tipo de financiación.
Hipótesis alternativa 2: la publicación de resultados depende del género del ensayo clínico.

Hipótesis alternativa 3: la publicación de resultados depende de la fase del ensayo clínico.
Hipótesis alternativa 4: la publicación de resultados depende de si es un estudio multicéntrico o no.
Todas la hipótesis alternativas son ciertas (p-value<0.05), la publicación de resultados depende de todos estos factores, pero ¿en qué medida?
El hallazgo de efectos estadísticamente significativos (cuando se rechaza la hipótesis nula) pueden ser irrelevantes cuando son de baja magnitud, lo que puede ocurrir cuando las muestras son bastante grandes. Debido a que tenemos muchos datos, es necesario estudiar el tamaño del efecto de cada hipótesis.
El tamaño del efecto es el siguiente:

Contraste de hipótesis	**Tamaño del efecto**	
Study.Results - Funded	0.279	Mediano
Study.Results - Gender	0.047	Pequeño
Study.Results - Phase	0.339	Mediano
Study.Results - Multicentric	0.19	Pequeño

Por lo que podemos decir que la publicación de resultados no depende o tiene un dependencia muy pequeña con el género y con la variable 'Multicentric'. También podemos decir que existe una dependencia mediana con el tipo de financiación y con la fase del estudio.

Cuando una de las variables tiene más de dos niveles y el test $\chi 2$ de independencia resulta significativo puede ser de interés estudiar en qué niveles se encuentran las diferencias significativas. Para ello haremos uso del análisis de los residuos de pearson. Si el valor residual estandarizado de un nivel supera el 1.95, se considera significativo.

Se genera un fichero de R para analizar estadísticamente los datos. Este código de R abre el fichero con los datos limpios y aplica los conceptos estadísticos comentados anteriormente.

5. Encontrar el hilo argumental a partir del análisis de los datos.

Con todos los datos seleccionados y filtramos ya empezamos a entrever cuál será el hilo argumental de nuestro reportaje de periodismo de datos.
Ya podemos decir que la gran mayoría de los ensayos clínicos no publican sus resultados en la web de clinical trials. Aproximadamente solo un 30% de ellos tienen resultados publicados en la web. Estos resultados parecen, a priori, independientes del tipo de enfermedad. Si hacemos una comparación rápida con otros estudios (con los mismos parámetros de búsqueda) que estudian otro tipo de enfermedad encontramos lo siguiente:

- Alzheimer, encontramos que solamente un 26.9% de los estudios publican sus resultados.
- Parkinson: un 41.5% publican resultados.
- Diabetes: un 35.5% publican resultados.

- Obesity: un 21.8% publican resultados.
- Asma: un 36.7% publican resultados.
- Radioterapia: un 23.8% publican resultados.

Aunque el porcentaje de publicación varía en cada caso, se podría decir con bastante seguridad que en general los porcentajes de publicación son inferiores al 50%. De esto podemos extraer que gran parte de la evidencia científica en la rama de la medicina no se hace pública.

En el caso de nuestro país, encontramos 693 ensayos clínicos registrados en el mismo periodo. De estos solamente 332 tienen resultados (47,9%).
Si comparamos este resultado con otros países, por ejemplo:

- Estados Unidos, de los 4003 registrados solamente 1745 tienen resultados (43,6%).
- UK: de 660, 296 tienen resultados (44,8%).
- Alemania: encontramos 350 (45.7%) con resultados de los 765 publicados.
- Francia: de 885, 327 tienen resultados (36.9%).
- Canadá: de 775, 339 tienen resultados (43.7%)
- China: de 452, 113 tienen resultados (25.0%).

Por lo que podemos decir (se necesita estudiarlo más a fondo) que el ratio de publicación parece que no depende de la cantidad de estudios registrados o del tamaño del país que los registre. Habrá que evaluar si esta afirmación es correcta y cuáles son exactamente los factores que influyen en la no publicación de resultados.

Queremos averiguar también cuánto influye el tipo de financiación: pública o privada. Por ejemplo, en la siguiente imagen podemos ver todos los ensayos con resultados publicados en la web (2362) según el tipo de financiación:

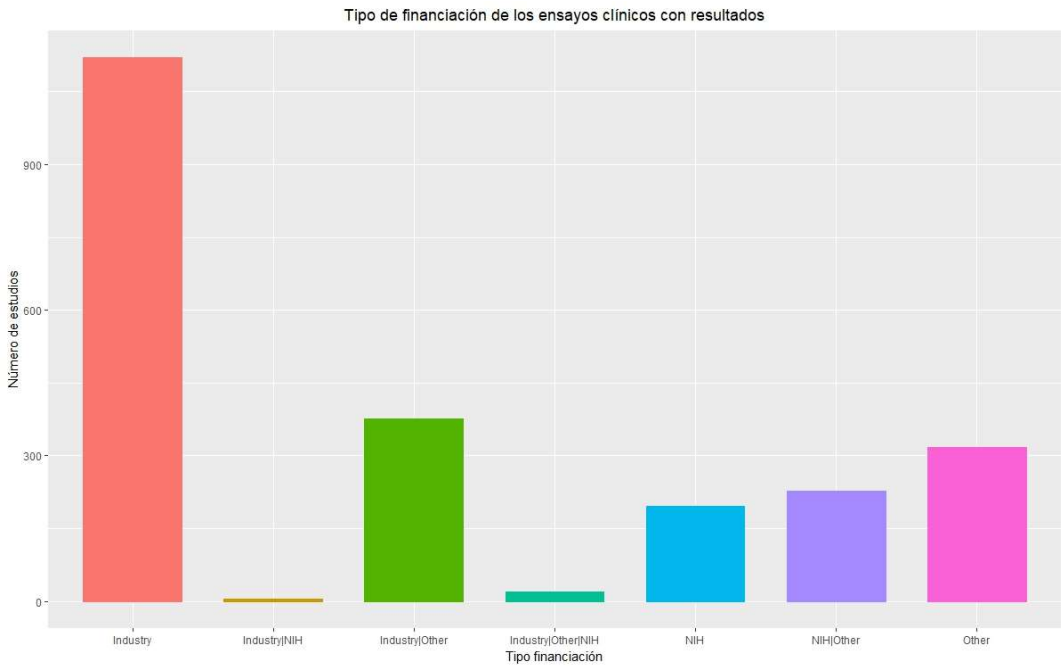

Imagen IX. Tipos de financiación

En la imagen se observa que la gran mayoría de los ensayos clínicos que presentan resultados están financiados por la industria (financiación privada, aproximadamente un 49.5%), seguidos de una combinación de pública y privada (Industry & Other, 16.7%).

Con respecto a si los ensayos clínicos cumplen la ley del 2007 de la FDA, podemos concluir que la gran mayoría de los ensayos clínicos en oncología no cumplen la ley, ya que solamente un 11% publicada sus resultados dentro de los 12 meses siguientes a la PCD.

Con respecto a los estudios españoles (se han seleccionado todos aquellos que incluyan un centro español, sean multicéntricos o no), observamos lo siguiente:

- De estos ensayos solamente publican resultados el 47.8% y de estos solamente un 8.1% cumple con la fecha de publicación.
- Si vemos la distribución de estudios registrados por provincia, un 93% son estudios multicéntricos. Para los estudios que no son multicéntricos, Madrid es la provincia que más estudios registra, seguida de Barcelona y Navarra.