3

Procesamiento de datos: paquetes estadísticos

Antonia Guillén Serra

3.1. Introducción

En el presente capitulo nos centraremos fundamentalmente en el proceso de gestión de datos, también conocido como "Clinical Data Management". En general se trata de un proceso complejo que engloba un conjunto de fases secuenciales cuyo principal objetivo es garantizar que las bases de datos provenientes de un estudio sean precisas, válidas, completas y seguras (Figura 1). Es importante no caer en el error de algunos investigadores al considerar esta fase de la investigación menos importante que las fases de diseño y análisis, dado que es la base a partir de la cual se desarrollará todo el proceso analítico y constituye el fundamento para extraer las conclusiones finales. La falta de calidad durante el proceso de gestión de datos puede afectar gravemente la validez interna y externa de un estudio cuyo diseño haya sido impecable. Si los datos analizados no se corresponden con la realidad, tampoco servirán de nada las herramientas estadísticas más potentes ni los programas informáticos de última generación o el personal más cualificado. Por tanto, debe tenerse en cuenta que el coste de adaptar medidas para proteger la calidad de los datos es muy inferior al que

representa poner en duda la validez de una investigación. Es por eso que desde la década de los 80 se ha ido desarrollando una línea de investigación entorno al tratamiento de los datos y al concepto de calidad de éstos (data quality). Además, no debe olvidarse que en este proceso intervienen investigadores y/o monitores, informáticos, grabadores de datos, administradores de bases de datos y estadísticos.

También se introducirá el concepto de depuración, que se define como un proceso a realizar necesariamente tras haber recogido los datos para detectar y corregir los errores que, por distintas vías, contiene nuestra base de datos.

Una de las aplicaciones del procesamiento de datos que más ha evolucionado ha sido el desarrollo de paquetes estadísticos. Conceptualmente un paquete estadístico es un programa informático específicamente diseñado para el procesamiento de datos con el objetivo de resolver problemas de estadística descriptiva o inferencial. El entorno legislativo que marca la normativa de las investigaciones indica que en el protocolo del estudio se debe identificar el sistema informático que se va a utilizar y éste debe ser un sistema trazable, seguro y fiable que asegure la confidencialidad de los datos, que permita identificar los usuarios que acceden al mismo y las modificaciones que se producen en su contenido a través del tiempo.

Con el desarrollo de la microinformática la utilización de los paquetes estadísticos ha pasado de los centros de investigación, empresas o instituciones públicas hasta poder instalarse en cualquier ordenador que en la actualidad se esté comercializando. Hoy por hoy, cualquier ordenador personal puede tener instalado alguno o varios paquetes estadísticos. Aunque las opciones que nos ofrecen sean cada vez más numerosas, esto no supone un obstáculo de complejidad para su utilización, lo que hace que el número de profesionales de enfermería que se interesan por ellos sea cada vez mayor. Conocer brevemente la utilidad y características de los diferentes paquetes estadísticos más utilizados en la investigación científica nos ayudará a elegir el más idóneo para la realización de nuestro trabajo.

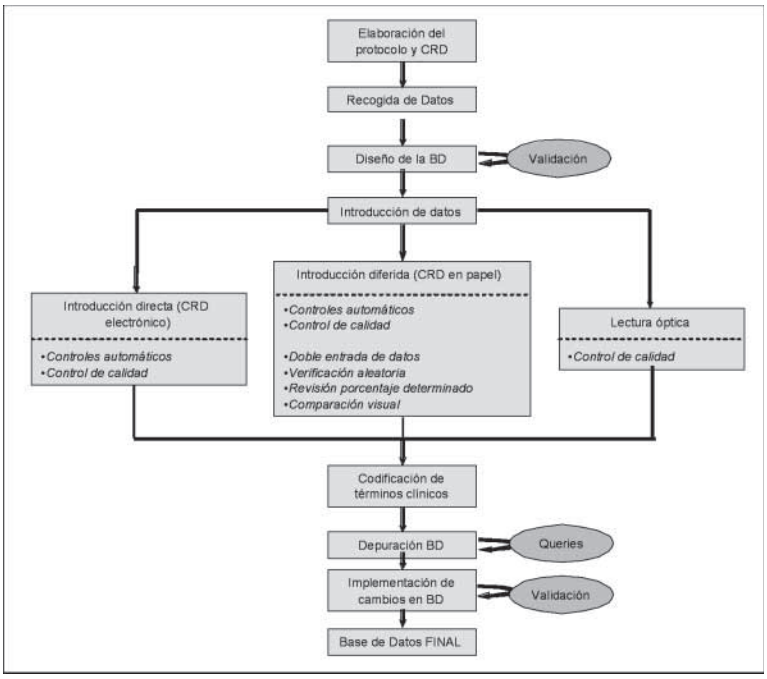

Figura 1. Esquema del proceso de Clinical Data Management

3.2. Gestión y calidad de los datos

Una vez se ha puesto en marcha la recogida de datos del estudio, ya podemos empezar a preparar la base de datos que contendrá toda la información que consideremos necesaria para realizar el análisis que hemos planteado inicialmente en el protocolo del estudio. Las tareas que requieren explorar un conjunto de datos pueden parecer un tanto mundanas e inconsecuentes, pero son una parte esencial para cualquier análisis y no deben ser olvidadas por los analistas o programadores. Por tanto, una buena gestión de datos tiene una relación directa con la minimización del coste económico y de tiempo a largo plazo, del mismo modo que determina que la presentación de los resultados obtenidos goce de una mayor credibilidad.

3.2.1. Diseño de la base de datos

En función del análisis estadístico a realizar y del diseño del cuaderno de recogida de datos (CRD), se determinará la estructura de la base de datos que incluye la definición única de las variables del estudio y su formato, junto con la

estandarización de las unidades de investigación, de manera que cada campo del CRD tenga una relación directa con una variable de la base de datos (BD).

Para la creación de la BD podemos utilizar diferentes programas informáticos que permitan tabulaciones de datos como pueden ser ACCESS, EXCEL, DBASE o se puede crear directamente la BD con el paquete estadístico que se vaya a utilizar. El único requisito es que las tablas de datos sean fácilmente exportables a otros formatos. Generalmente, las variables (edad, TAS, TAD) se representan en columnas y las unidades de investigación (pacientes, pruebas), en filas.

En la elaboración de cualquier BD es muy importante conocer los requisitos para hacer más fácil la introducción de los cuestionarios y el análisis que se va a realizar, siendo un trabajo que facilitará la programación y/o ejecución de las pruebas estadísticas y su posterior interpretación. Una forma de asegurarnos que la BD está preparada para realizar la introducción de los casos del estudio, es introducir casos imaginarios que permitan detectar errores de formato o posibles omisiones de variables. A este proceso se le denomina Validación del diseño de la BD. A continuación, se citan los parámetros más importantes en la creación de variables.

- Definición de las variables: Cada variable debe tener un nombre único, a poder ser identificativo de su contenido y que normalmente no debe exceder de 8 caracteres (Ej. EDAD).
- Formato de la variable: Definiremos el formato de la variable en función del contenido de ésta, de manera que podrán ser clasificadas en: numérica, alfanumérica, hora, fecha, día de la semana, etc. En caso de variables numéricas con decimales, fijaremos el número máximo de dígitos enteros de la variable y el número máximo de dígitos para los decimales.
- Etiquetas de las categorías: Las variables categóricas tienen asignado un valor para cada categoría. Para cada una de estas categorías podemos describir los valores que recoge asignando una etiqueta a cada categoría (1='Hombre', 2='Mujer').
- Etiqueta de la variable: Opcionalmente, definiremos una etiqueta para cada variable con la descripción de su contenido. En caso de que el nombre de la variable sea poco clarificador, su etiqueta puede sernos útil durante la introducción de datos.
- Valores perdidos de la variable (missing): Definiremos el valor o valores de cada una de las variables que representen la falta de información, valores perdidos o datos completados parcial o incorrectamente. La existencia de valores desconocidos es una constante en cualquier investigación de difícil control aunque estos valores se ven disminuidos al aumentar la

rigurosidad con la que se haya planificado la recogida de datos. La proporción de valores desconocidos constituye un indicador más de la calidad del proceso de recogida de datos y de la información que se analiza, por lo que consideraremos más importante el tratamiento que se va a dar a estos valores en el protocolo, así como durante la recogida de datos.

El tema referente a los missings es muy controvertido puesto que presentan un problema en los análisis que debemos realizar. Además de la falta de información que esto supone, como consecuencia hay una pérdida de muestra y de potencia estadística, ya que muchos de los paquetes estadísticos eliminan automáticamente las observaciones que tienen valores missing para cualquiera de las variables usadas en el análisis. Es por eso que deberemos intentar recuperar estos valores, tal como veremos en el apartado 3.2.4. Depuración de la Base de Datos.

Para poder diferenciar entre los valores que realmente no podremos recuperar y aquellos que son susceptibles de solucionar, distinguiremos en la base de datos las anotaciones que pueden encontrarse en el CRD mediante códigos asignados a tal efecto. Si la pregunta formulada está específicamente contestada con una de las respuestas que aparecen en la tabla siguiente (Tabla 1), se le asignará un código especial. Normalmente, suelen ser códigos como 9, 99 ó -1, siempre y cuando estos valores no se solapen con otros valores de la variable y no se caiga en el error de que estos valores sean incluidos en el análisis.

RESPUESTAS / ANOTACIONES	SIGNIFICADO	SOLUCIÓN
Campo en blanco	No se ha determinado el valor de la respuesta	Solucionable
Dato no legible	Se ha determinado la respuesta pero no se consigue descifrar el significado	
No disponible, Nor avaible, No recogido, No determinado (ND)	No se dispone de la información	Imposible de recuperar
No realizado (NR, ND), No hecho, Not done	Prueba no formulada o pregunta no realizada	
No conocido (NC), desconocido (DESC), Unknow, (UNK, NK)	Se desconoce la respuesta o el valor	
No aplicable (NA)	No procede aplicar la pregunta	

Tabla 1. Valores considerados missing, significado y posibilidad de solución

Así pues, un campo en blanco lo consideramos missing e intentaremos recuperarlo siempre que sea posible. Sin embargo, un campo «No disponible» o «No determinado» se considera un missing con ligeras diferencias: el investigador o persona encargada de la recogida de datos nos avisa de que ese campo no va a ser recuperable, por mucho que lo intentemos. Es una forma de asegurarse, especialmente en los ensayos clínicos, de que todas las respuestas posibles están recogidas en la BD. Posteriormente, todas estas respuestas serán interpretadas como missing.

De cara al análisis, existen diversas técnicas de imputación de datos en cuanto a los datos faltantes o missing. El método más utilizado es el de despreciarlos, de forma que éstos no aparecen en el análisis. Sin embargo, existen otras técnicas de imputación de valores que permiten sustituir los valores faltantes por: la media de la variable; en caso de un estudio longitudinal, la media de las observaciones anterior y posterior en el tiempo o bien la imputación mediante la técnica conocida por LOCF (Last Observation Carried Forward) que consiste en asignar el último valor conocido. Por ejemplo: en un estudio longitudinal de insuficiencia renal se pretende estudiar la evolución de los valores de hemoglobina en sangre a los 2 y 4 meses de seguimiento; en caso de no tener la observación final, podríamos "arrastrar" el último valor conocido (2 meses) y considerar éste como el valor real. Otros métodos más complejos para la imputación de los valores missing están relacionados con imputaciones múltiples a través de varias variables. Como vemos, es un tema que admite multitud de opciones, siempre y cuando las decisiones se tomen con cierto criterio.

3.2.2. Introducción de datos

Las estrategias utilizadas para mejorar la calidad de los datos durante su grabación tienen el objetivo de garantizar que la información esté exenta de inconsistencias, con el menor número de valores desconocidos y que el tiempo requerido se ajuste al calendario del estudio.

Actualmente los métodos mas utilizados son: la introducción diferida, que es la metodología más utilizada hasta el momento donde los datos se encuentran en formato papel y se graban en soporte magnético, y la introducción directa, donde es el propio investigador el que realiza la grabación en la BD mediante lo que es conocido como CRD electrónico. Existe otro método de introducción más novedoso que se realiza mediante lectura óptica. Esta técnica escanea las fuentes documentales originales simplificando la grabación de los datos y mejorando en gran medida su calidad; si bien permite evitar los errores introducidos por el operador y disminuye el coste de la introducción, a esta técnica aun le quedan algunos problemas por solventar. Algunos de estos inconvenientes que

presenta son: falta de formación y de personal experimentado, utilización de un software poco fiable y deficiente tratamiento de variables con formato abierto. En la actualidad, se están desarrollando diferentes técnicas para poder reconocer los caracteres escritos y establecer protocolos fiables que solventen los problemas de legibilidad e interpretación de los datos.

Para validar los datos introducidos se pueden implementar diferentes tipos de controles de calidad. Los controles automáticos se aplican tanto en la introducción directa (CRD electrónico) como diferida (CRD en papel) y funcionan programando filtros interactivos que impedirán que los valores inconsistentes sean grabados. Cuando se realiza la introducción de forma directa, programando todos los filtros posibles a priori, tenemos la ventaja de poder contrastar el error y, si es preciso, recuperarlo al momento. En el caso de utilizar el método de introducción diferido destacaremos los siguientes controles de calidad:

- *La doble introducción de datos:* La información de los cuestionarios es introducida por duplicado en dos bases de datos preferiblemente por personas distintas para evitar posibles interpretaciones de los datos y, una vez introducidos todos los casos, son comparadas para detectar aquellos valores que no coinciden en los dos registros. Una vez detectados los errores se corrigen los datos incorrectos revisando la información original. La eficacia de esta técnica está demostrada por algunos autores reduciendo la proporción de los errores hasta un 30% (Bonillo.A), pero tiene sus limitaciones: por un lado, la repercusión en el coste del estudio que supone la doble introducción, y por otro, el requerimiento de un perfil de los operadores con experiencia y familiarizados con el uso de formularios utilizados en el estudio.
- *Verificación de variables más importantes:* Consiste en grabar dos veces las variables imprescindibles del estudio y una sola vez el resto de la información. Los datos son introducidos por el mismo operador con el suficiente retraso para evitar el efecto recuerdo. Este proceso tiene la ventaja de reducir los costes y de proporcionar un feed-back de los errores que se van cometiendo.
- *Revisión de un porcentaje determinado de casos:* Una vez informatizados los datos se revisan un porcentaje determinado de los casos introducidos para estimar el tanto por ciento de errores cometidos en la introducción. Si el porcentaje de error es inferior al fijado previamente, dependiendo del tipo de estudio y tipo de variables utilizadas, daremos por válida la BD. Si por el contrario, el porcentaje de error es excesivamente elevado tendríamos que seguir revisando la BD o reintroducir los datos.
- *Comparación visual:* Una o dos personas comparan visualmente los datos

grabados con los registrados en las fuentes originales. Constituye un procedimiento costoso, sólo recomendable en estudios de pocos casos y sin opción a la utilización de otros procedimientos.

Además de utilizar uno de estos procesos de control de calidad de la BD, también deben revisarse los valores extremos o anormales de las variables con la finalidad de detectar errores de introducción. Por otro lado es conveniente cruzar las variables de la BD que nos puedan indicar un error.

3.2.3. Codificación de términos médicos de la base de datos

La codificación es la fase dentro del proceso de gestión de datos que requiere de más conocimientos médicos y farmacológicos. En los estudios clínicos, el origen habitual de estos datos provienen de la historia clínica de los pacientes y nos podemos encontrar con una gran parte de información que no necesita ningún tipo de codificación como pueden ser variables numéricas (edad, TAS, valores analíticos, FC, etc..) o variables cualitativas o categóricas (gravedad del asma, hábito tabáquico, etc). Sin embargo, existen otras variables que contienen habitualmente información escrita o alfanumérica como son: enfermedades concomitantes, reacciones adversas y medicación concomitante, y que en muchos casos, son el objetivo principal de nuestros estudios y que requieren de codificación para ser resumidas o analizadas.

A pesar de que un mismo término puede transcribirse de formas muy distintas, nos encontramos que tienen un mismo significado. Por ejemplo, como acontecimientos adversos, los "picores" y el "prurito" podrían agruparse en un mismo término y, en el caso de medicación, el principio activo de un medicamento específico puede ser fabricado por distintos laboratorios farmacéuticos y recibir diferentes nombres comerciales (Gelocatil® y Termalgin®). Este problema aumenta con el uso de contracciones para los nombres compuestos como (HTA, IRC, DM, ARAII, IECA...), además de la cumplimentación de los cuestionarios con letra en ocasiones poco clara. Para evitar estos inconvenientes, en algunos estudios se ha intentado que el investigador registrara la información de forma codificada, pero es un procedimiento que no cuenta con gran aceptación entre algunos profesionales.

Por tanto, el objetivo de la codificación consiste en unificar todos estos conceptos para poder ser tratados estadísticamente durante el análisis. Además, la codificación es una buena oportunidad para revisar los cuadernos de recogida de datos y observar irregularidades fácilmente detectables.

Existen bases de datos estandarizadas tanto nacionales como internacionales que son publicadas y actualizadas periódicamente, tanto en soporte papel como

en soporte magnético. Este tipo de codificación es imprescindible en los ensayos clínicos y forma parte de las herramientas de los distintos departamentos de farmacovigilancia. A continuación citamos algunas de ellas:

- MedDRA (Medical Dictionary for Regulatory Activities): Es el diccionario imprescindible de codificación, comunicación y registro de las reacciones adversas (RA) en los departamentos de farmacovigilancia. Tiene una terminología médica internacionalmente aceptada y clínicamente validada, con versiones en ocho idiomas distintos para uso en todos los procedimientos de regulación de medicamentos en los que deban utilizarse términos médicos. Esta BD ha sido desarrollada bajo la normativa de las ICH (International Conference on Harmonisation) y aparece ante la necesidad de regular y establecer un estándar único e internacional para la regulación de medicamentos, que satisfaga los requerimientos tanto del registro como de las posteriores tareas de farmacovigilancia. MedDRA incluye términos que provienen de los diccionarios COSTART (Coding Symbols for a Thesaurus of Adverse Reaction Terms- Preferred Terms and Glossary Terms 5° edición), WHO-ART y su adaptación Japonesa (J-ART, 1996), HARTS (Hoechst Adverse Reaction terminology 2.2) y una gran parte de los términos que integran la Clasificación ICD9 Y ICD9-MC y la adaptación Japonesa del ICD9 (MEDIS). MedDRA es la terminología médica adoptada por Europa, Japón, Estados Unidos y Canadá.
- ICD-9-MC: Modificación clínica de la Clasificación Internacional de Enfermedades 9° Revisión (CIE-9) de la Organización Mundial de la Salud. El término clínico se utiliza para subrayar el propósito de la modificación, prefiriéndola algunos autores, a pesar de que existen versiones más novedosas. Se utiliza en el campo de las clasificaciones de datos referentes a morbilidad, ordenación de historias clínicas, revisiones de cuidados médicos y programas de cuidados ambulatorios. También es requerida en las estadísticas básicas de salud, descripción de cuadros clínicos o codificaciones que precisan ser más específicas que aquellas que se necesitan exclusivamente para agrupaciones estadísticas y análisis de tendencias.
- Clasificación ATC de sustancias farmacéuticas para uso humano: Sistema europeo de codificación de sustancias farmacéuticas y medicamentos estructurados en cinco niveles con arreglo al sistema u órgano afectado, efecto farmacológico, indicaciones terapéuticas y estructura química del fármaco. A cada fármaco le corresponde un código ATC y este se especifica en su ficha técnica.

- WHO-DRUG Dictionary: Diccionario de medicamentos de la OMS que se basa en la clasificación ATC. Se considera una herramienta esencial para la codificación de medicamentos de ensayos clínicos.
- WHO-ART Dictionary (WHO Adverse Reaction Terminology): Terminología de la OMS para reacciones adversas causadas por medicamentos.

Existen otros parámetros como los resultados de laboratorio, tipos de pruebas de diagnóstico, especificaciones de profesiones de riesgo, etc. que antes de ser analizados deben codificarse para ser informatizados. Aunque el volumen del estudio sea pequeño y los análisis que pretendemos realizar sencillos como para una tabulación manual, igualmente será necesario codificar utilizando un manual de codificación especialmente diseñado para el mismo.

3.2.4. Depuración de la base de datos

Después de la introducción de los cuestionarios de recogida de datos se debe realizar el proceso de depuración y limpieza de los mismos, con el objetivo de garantizar que el conjunto de datos que se someten a la etapa de creación de nuevas variables y análisis estadístico, contienen el menor número de valores desconocidos y no incluyen inconsistencias. Al ser manipulaciones que se efectúan sobre la BD, tanto la detección de los errores como la corrección de los mismos deben ser registrados y documentados convenientemente.

En esta etapa se evalúa la validez y la utilidad de los datos introducidos. Lo ideal seria iniciar este proceso cuando llegan los primeros resultados como si de un estudio piloto se tratara, ya que la detección temprana de los errores ayuda en gran medida a minimizarlos y corregirlos en un futuro. Los principales controles para la detección de errores serian los siguientes:

- *Identificación del CRD:* El código del paciente debe ser único, sin repetición, de forma que los registros estén unívocamente identificados.
- *Consistencia básica del conjunto de datos:* Se aplicarán filtros y chequeos lógicos y estructurales sobre las variables que evaluarán la consistencia de los datos introducidos. Se asegura que el dato es consistente respecto a su campo o respecto a otros valores. Por ejemplo, una prueba de detección de enfermedad prostática (PSA) nunca puede ser realizada a mujeres.
- *Consistencia estadística del conjunto de datos:* La aplicación de determinadas pruebas estadísticas que pueden detectar falta de consistencia en el conjunto de datos, por ejemplo: valores atípicos, respuestas aberrantes o valores dentro del dominio pero extremos en su grupo de pertenencia. Para que un valor atípico lo sea, no debe tener sentido sustancial. Por

ejemplo, un potasio de 8 mEq/l es un dato fuera de rango; sin embargo este valor puede no ser un valor estadísticamente atípico en un pequeño grupo de pacientes con una insuficiencia renal crónica en tratamiento sustitutivo, cuyo rango esperado puede estar entre 3.5 y 5 mEq/l.

- *Valores extremos:* Los valores extremos para una variable se denominan "outliers" (valores atípicos).
- *Datos incompletos:* Los datos incompletos son valores que faltan o completados parcial o incorrectamente. Como hemos definido en el apartado anterior, son los valores también denominados missing.

Los datos que no cumplan los requisitos serán revisados y corregidos mediante el proceso de generación de queries. El término inglés *query* significa "pregunta, duda". Esta palabra hace referencia a las preguntas que se realizan a los responsables de la recopilación de los datos del estudio (investigadores, monitores del estudio y/o ayudantes) para resolver dudas o inconsistencias generadas en el proceso de depuración y validación de la BD.

La necesidad de una *query* puede ser identificada durante la monitorización (revisión de los cuadernos), el proceso de entrada de datos, validación o codificación de los mismos. La *query* será generada y la respuesta actualizada en la BD cuando se disponga de dicha información. Cualquier cambio en la BD quedará registrado. Estas *queries* también deben ser archivadas y documentadas específicamente, puesto que son modificaciones realizadas en la BD y por tanto indispensables para seguir el principio de trazabilidad que hemos nombrado al principio de este capítulo.

Ante la necesidad de contar con herramientas que permitan chequear la consistencia de los datos capturados, algunos paquetes estadísticos han incorporado en sus últimas versiones instrucciones dedicadas a este fin.

Una vez finalizada la corrección de las incidencias debe efectuarse una valoración de la calidad de los datos obteniendo el porcentaje de error por sujeto y por variable. Otro aspecto de la calidad es el porcentaje final de los valores desconocidos por sujeto y por variable que no han podido recuperarse. Una vez que los datos estén depurados, la BD será cerrada, asegurando que los datos no sufrirán ninguna manipulación posterior.

3.3. Herramientas para el análisis de datos

Tanto para el proceso de introducción de datos como en la depuración y proceso de Clinical Data Management en general, no es fácil elegir las herramientas necesarias para llevar a cabo estas funciones. Muchas veces, será una

combinación de todas ellas, puesto que cada una se adecuará por distintos motivos a nuestras preferencias. A continuación, se intentará describir los principales paquetes estadísticos y su utilidad.

3.3.1. Definición de paquetes estadísticos

Un paquete estadístico es un programa informático de cálculo de análisis estadístico que incluye frecuentemente la confección de gráficos para tener una interpretación más visual de los resultados. Generalmente suelen tener una ventana de visualización de los datos en formato de panel de celdas (filas y columnas) y una ventana de los resultados obtenidos y otra ventana de programación propia de cada software.

Su utilización se dirige preferentemente a la realización de complejos y costosos cálculos que implica el desarrollo de la estadística actual. De esta forma, procesos de análisis matemático que antes eran interminables y que podían ocupar períodos de tiempo extensos se resuelven ahora en períodos relativamente cortos.

En la actualidad existen muchos y variados paquetes estadísticos en el mercado, desde los más simples, que sólo incluyen la estadística descriptiva, hasta los más complejos que realizan todo tipo de cálculos, incluso algunos de ellos se han especializado en concreto para el desarrollo de las técnicas estadísticas más avanzadas. Para manejarlos se puede hacer en modo de comando, es decir, a través de órdenes escritas, o por menús, facilitando más su manejo.

3.3.2. Aspectos básicos y características de evaluación de los paquetes estadísticos

Por tanto, saber cuál es el paquete estadístico idóneo para cada usuario dependerá, no tan sólo de la aplicación que se le quiera dar y del proyecto al cual va dirigido, sino que además hay que tener en cuenta una serie de requisitos imprescindibles para su adquisición, instalación y utilización. Los aspectos básicos a tener en cuenta a la hora de adquirir o utilizar un paquete estadístico pueden ser englobados en dos tipos: por un lado, las características imprescindibles que debe tener el usuario, y por otro, las características propias de la herramienta en si.

3.3.2.1. Características del usuario

En nuestra opinión, no siempre es necesaria la utilización de programas punteros en el campo de la estadística. Para casos concretos se pueden encontrar

hojas de cálculo o bien pequeños programas de elaboración propia que pueden ser la solución a nuestros problemas dado que pueden efectuar cualquier prueba que necesitemos, por muy extraña que sea. En este caso, la dificultad se centra en discernir entre los programas de calidad y aquellos que no funcionan correctamente. Los profesionales que sólo se aproximan de forma puntual a problemas estadísticos y que buscan soluciones poco sofisticadas y puntuales, deberían en primer lugar reexaminar las rutinas estadísticas de cualquier hoja de cálculo, incluyendo tal vez algunas de las macros de libre distribución que se puedan encontrar en Internet.

Usando una combinación con otro tipo de programas que extienden su capacidad gráfica, tales como hojas de cálculo, procesadores de texto y/o programas específicos para la representación de datos, se pueden realizar análisis más sofisticados que los que precisan la mayoría de estos usuarios con un coste nulo y sólo los más expertos encontrarían a faltar alguna de las herramientas adicionales que incluye su versión comercial. Los requerimientos necesarios para su elección en cuanto a nuestras características particulares son:

- *Requisitos del hardware:* Debemos comprobar si el hardware que poseemos es adecuado para la utilización de un paquete estadístico concreto, como son: memoria RAM, disco duro, tarjeta gráfica, etc.
- *Conocimientos del usuario:* Es aconsejable que el usuario tenga unos mínimos conocimientos de informática y programación, de lo contrario tendrá que renunciar a realizar por su cuenta cierto tipo de análisis que vayan más allá de los comunes, además de poseer conocimientos básicos en estadística. En muchos casos, no tener conocimientos amplios en estadística puede ser solventado por la ayuda que estos paquetes ofrecen (manuales o ayudas interactivas), aunque en algunos casos puede resultar algo peligroso dado que se pueden aplicar pruebas sin conocer aspectos importantes como las condiciones de aplicación o la interpretación de los resultados.

3.3.2.2. Características generales del paquete estadístico

La elección del programa que vamos a utilizar debe ajustarse tanto a nuestras necesidades como a nuestras posibilidades. Estas son las características que destacamos:

- *Coste:* Empresas y centros de cálculo y/o estudio donde existan departamentos de estadística que necesitan manipular grandes bases de datos aparte de las soluciones antes mencionadas, deberían plantearse hasta qué punto les es más provechoso adquirir una licencia para usar uno de los

grandes paquetes estadísticos. Existen en el mercado programas cuyas licencias anuales tienen un coste que están fuera del alcance de la mayor parte de los usuarios, como por ejemplo SAS o SPSS; sin embargo, también disponemos de programas de libre distribución en Internet o con un coste más asequible.

- *Facilidad de programación:* Se debe valorar el modo en que opera el programa, facilidad de manejo, funcionamiento con menús, software interactivo, capacidad de importación y exportación de BBDD, complejidad de lenguaje de programación, adecuación de las ayudas y/o tutoriales, entre otros.
- *Potencia computacional:* En función de la complejidad del análisis a realizar, debe asegurarse que el paquete estadístico elegido sea capaz de efectuarlos, adaptándose al modo en que nosotros queremos trabajar: rapidez de ejecución, simplicidad de programación, confección de gráficos de forma fácil y resultados entendibles.
- *Tamaño de la base de datos:* La mayoría de los paquetes estadísticos pueden manejar conjuntos de datos de un volumen considerable sin problemas.

En este libro hemos valorado la utilización del paquete estadístico SPSS para los ejemplos, por ser uno de los más utilizados en las publicaciones científicas de Enfermería Nefrológica dispone de un amplio conjunto de métodos estadísticos (multivariados, regresión logística y análisis de supervivencia) y tener un sistema de programación mediante menús que resulta muy sencilla.

3.3.3. Paquetes estadísticos más difundidos

A continuación se citan los paquetes estadísticos más difundidos en el ámbito sanitario, junto con las direcciones de Internet donde obtener información específica de cada uno de ellos.

- **SAS, Statistical Analysis System** (SAS Institute Inc., Cary, NC) *http://www.sas.com.* Paquete estadístico que ofrece numerosas posibilidades en cuanto a análisis estadístico y gestión de bases de datos. Requiere de conocimientos bastante elevados en programación. Elevado coste asociado a su licencia.
- **SPSS, Statistical Package for the Social Sciencies** (SPSS Inc., Chicago) *http://www.spss.com.* Programa que permite realizar fácilmente análisis estadísticos desde los más simples hasta un nivel elevado de sofisticación, con fácil programación mediante menús y cuadros de diálogo. Buena importación y exportación de ficheros. Dispone de manuales y ayudas. Es uno de lo paquetes más popularizados en todos los sectores gracias a su facilidad de manejo.

- **Epi Info** (Center for Diseases Control/WHO, Dean et al). *http://www.cdc.gov/epiinfo.* Programa de dominio público para la creación de bases de datos y análisis estadístico, especialmente útil en el ámbito de la epidemiología. Representación de gráficos y mapas.
- **Minitab Statistical Software** (State College, Pa) *http://www.minitab.com.* Programa de fácil manejo para la realización de análisis estadísticos con buenos gráficos. Muy usado tanto por estudiantes como por profesionales del sector. Requiere licencia.
- **Stata Statistical Package** (Stata Corporation, Computing Resource Center, College Station, Texas) *http://www.stata.com.* Paquete estadístico gratuito muy utilizado por investigadores médicos, bioestadísticos, epidemiólogos, con métodos estadísticos muy potentes y gráficos asociados.
- **Microsoft Excel** *http://www.microsoft.com/office/excel/default.htm.* Aunque no se trata de un paquete estadístico dispone de funciones y macros disponibles en Internet para pocos datos y análisis sencillos, sin necesidad de recurrir a otros más complejos y costosos. Recomendado para análisis puntuales.
- **Statcrunch** *http://www.statcrunch.com.* Programa gratuito de libre acceso bastante utilizado para iniciarse en la estadística. Las capacidades de este software se han ido incrementando así como también el número de usuarios.

3.4. Consideraciones importantes

Como se ha visto en este capítulo, el proceso de gestión de datos que hemos expuesto es metódico y un tanto complejo, y abarca una serie de tareas que incluyen desde la entrada de datos a la codificación médica, entre otros. Dado que es quizá la fase más dinámica del proceso y en la que intervienen distintos profesionales, a pesar de que se ha representado una posible forma de trabajo, existen otras propuestas con variaciones en los nombres y ubicaciones de las fases. La bibliografía consultada y la práctica habitual recomiendan aplicar un proceso más o menos riguroso, sin olvidar que los datos que van a ser analizados deben haber sido recogidos, revisados, codificados y validados.

La calidad de los datos puede verse alterada en cualquier punto del proceso en que se obtienen o se modifican los datos; por eso, somos conscientes de que para mantener esta calidad se requieren una serie de actividades que, en ocasiones, son consideradas banales y difíciles de motivar, pero a la vez, imprescindibles. Resaltaremos de nuevo la importancia de esta fase de la investigación que

debe poder responder de forma precisa y oportuna a las preguntas que se puedan plantear en relación a la calidad de los datos, sobretodo ante el aumento de la preocupación por la mala conducta científica y el fraude en las ciencias biomédicas.

Los investigadores o personas responsables del estudio tienen el compromiso de mantener la documentación del estudio durante varios años después de su publicación y muchos autores consideran imprescindible un registro de auditoria, que es el mecanismo esencial para documentar e identificar los cambios que son realizados en los datos a cada paso: quién, dónde, cuándo y cómo han sido realizados.

Consideraremos por tanto que: una buena planificación del protocolo y objetivos del estudio, donde exista la elaboración de un CRD claro y simple que permita la creación de una BD acorde al mismo con pocos campos abiertos, además de una perfecta coordinación de todos los profesionales implicados en el proyecto y la utilización de una metodología pulcra en todas sus fases será imprescindible para el buen desarrollo de un estudio.