1. **Análisis de las variables del registro e identificación de los requisitos de información según los criterios normalizados para los registros de pacientes.**

Se ha realizado un estudio exhaustivo sobre las variables de interés que se estaban registrando hasta el momento en el anterior sistema software contratado por la Asociación Española de Nefrología Pediátrica. Como requisitos de información se han determinado, entre otros, los distintos tipos de datos existentes para, posteriormente, realizar las conversiones oportunas. De esta forma se ha diferenciado entre valores tipo fecha, numérico, texto, selección única y selección múltiple. Además se han establecido los requisitos necesarios para cada una de las variables contempladas anteriormente, definiendo valores límite, rangos mínimos y/o máximos, así como intervalos de normalidad que permitan una gestión más sencilla y cómoda para el profesional encargado de trasladar la información a la nueva plataforma.

1. **Diseño e implementación de una infraestructura de datos basada en el motor de bases de datos MySQL. Migración de los datos. Control de calidad de la nueva base.**

Se ha diseñado un procedimiento automático de migración capaz de transformar los datos de origen, desde una base de datos SQL Server, hasta el nuevo sistema de gestión de base de datos, que se utilizará a partir de la finalización de este proyecto, MySQL, adaptando éstos procedimientos para que satisfagan los requisitos identificados en el punto anterior.

Durante la ejecución de esta tarea se han detectado una serie de incoherencias existentes en los datos de origen, que ha dado lugar a la realización de un estudio pormenorizado de algunas variables implicadas en dichos conflictos, según las reglas de calidad de datos establecidas. Esto ha provocado que, posteriormente a la importación de los datos, se haya tenido que implementar un proceso de limpieza y control de los datos existentes en el registro, para poder eliminar dichas incoherencias y conseguir un punto de partida más acorde a la realidad clínica registrada de esos pacientes.

Esta limpieza y corrección de información ha provocado un retraso considerable en el cronograma inicial de tareas, puesto que en ocasiones se ha tenido que realizar esta tarea de manera manual, dada la imposibilidad de automatizarla. Se ha generado por tanto un procedimiento ad-hoc de verificación, corrección y control de calidad adicional al que se preveía inicialmente. Aunque este proceso haya provocado el correspondiente retraso, respecto a las estimaciones iniciales de duración del proyecto, es vital sacrificar ese tiempo dado que ha permitido garantizar la integridad de la información, que se venía insertando hasta el momento, así como evitar posibles errores futuros relacionados con dicha incidencia.

1. **Transformación de los datos mediante adaptación de las variables a los requisitos normalizados. Implementación del sistema de codificación diagnóstica de enfermedades renales según los criterios europeos del registro ERA-EDTA traducidos al español.**

Se ha aprovechado la importación de este registro a una nueva herramienta para dotar de nueva funcionalidad al sistema de entrada de datos. Ahora se realiza una clasificación diagnóstica en base a los estándares europeos existentes para ello y, además, se informa en ciertas variables de las salidas del rango normalizado de los valores introducidos. De esta forma se contribuye a mejorar tanto el proceso de inserción de datos, con mayor control sobre los mismos, como a realizar, en análisis futuros, una clasificación de los datos en base a dichos criterios estándares.

1. **Diseño de un cuaderno de recogida de datos electrónico según el software gratuito REDCap.**

Para adaptar el registro existente al sistema REDCap, se ha establecido una estructura equivalente a lo que sería un estudio longitudinal en el que cada paciente, a lo largo del tiempo, va generando distintos eventos de evolución registrados anualmente.

Se ha definido por tanto un evento inicial, de registro de paciente, otro para las distintas visitas de evolución y un último evento en el que se puedan añadir cualquier cambio de interés, asociado a cada uno de los sujetos contemplados en este registro, esto es, cambio de centro, entrada en programa IRT, recuperación de función renal, pérdida de función renal, perdida de seguimiento del paciente, etc.

También se ha generado un usuario específico para cada uno de los centros que actualmente están comprometidos en la introducción de datos en el registro. Se ha realizado la asignación de éstos pacientes a sus centros de referencia. Se está preparando todo para que el acceso a la herramienta no sea un hándicap a la hora de continuar con la colaboración en la toma de datos.

1. **Etapa de preprocesamiento para el análisis de datos.**

Se ha realizado una serie de tareas de preprocesamiento de datos con el objetivo de que posteriormente podamos llevar a cabo de la manera más efectiva posible un análisis y modelado predictivo de los datos. A partir de esta etapa de preprocesamiento se han detectado algunas inconsistencias en los datos, que han surgido producto al proceso de migración de la base de datos desde el sistema antiguo y que ya se ha comentado anteriormente en el punto 3. Estas inconsistencias detectadas a favorecido el refinamiento de la base de datos que actualmente se encuentra en explotación. Entre los procesos de limpieza de datos que hemos desarrollado se encuentran:

1- Detección de outliers: Se ha validado que los datos se encuentren en los rangos permitidos, y aquello valores anómalos han sido eliminados y substituidos teniendo en cuenta el método de imputación que más se adecua al tipo de variable (por ejemplo, la moda en datos categóricos, la mediana en datos numéricos, etc).

2- Valores omitidos: A partir de este proceso se han detectado elevadas cantidades de valores omitidos tanto en variables primarias como secundarias; estos valores perdidos se han transferido desde el sistema antiguo donde se evidencia que no existían reglas robustas de entrada de datos. A partir de este análisis se han podido implementar planes de contigencia para la implementación de reglas de entrada de datos en el nuevo sistema, además de la introducción de variables de cálculo automático que permitirán disminuir significativamente los errores de entrada de datos que posiblemente los especialistas puedan cometer. Por último, en aquellas variables en el que los niveles de valores perdidos no eran muy elevados se pudo realizar un proceso de estimación de dichos valores.

3- Normalización y estandarización de los datos: Para el correcto análisis predictivo que se realizará próximamente, se han realizado una serie de pasos que permiten estandarizar las variables a una misma escala, eliminar variables que no presentan varianza o ésta es casi nula y que por lo tanto no aportan información discriminativa para tareas predictivas, así como se ha tratado de eliminar la posible heterocedasticidad en variables numéricas.