PhUSE EU Connect 2018

Paper RW01

The Salford Lung Study: Challenges and Complexities with the Reporting of Electronic Medical Record Data in a Pioneering and Novel Approach to Clinical Trials

Daniel Kwok, GlaxoSmithKline, Uxbridge, United Kingdom
Jennifer Colthart, GlaxoSmithKline, Uxbridge, United Kingdom

ABSTRACT

The Salford Lung Studies (SLS) are ground-breaking trials designed to study the effectiveness and safety of initiating treatment with Relvar® Ellipta® in patients with Asthma and COPD versus existing continuing maintenance therapy. The usual way of investigating new medicine is to use a double blind, randomised controlled trial (RCT) – the gold standard for clinical research. This often has very strict inclusion/exclusion criteria, and while it is scientifically robust, it does not demonstrate how medicines may behave in a usual real-life setting. The aim of SLS was to keep as close to everyday practice as possible. Electronic medical record (EMR) data within the Salford region in the UK is shared between primary and secondary care in real time and enabled sites to monitor the patients carefully without being too intrusive.

We explore the challenges of collecting and analysing such data along with the complexities of working with Case Report Form (CRF) and EMR data in programming.

INTRODUCTION

Traditional Phase III double-blind randomised controlled efficacy trials tend to be highly controlled and select a patient population using strict inclusion and exclusion criteria. Patients are required to attend often relatively frequent scheduled visits and the investigator would often fill out a comprehensive CRF which has been specifically designed to capture data in a standard format for the study.

The statistical basis for such trials are very robust. However, there is a greater demand for data in a more “real-world” setting to complement pivotal Phase III studies, particularly from payers who make reimbursement and policy decisions that impact health care systems to evaluate drug effectiveness, safety and risk/benefit. Setting up and collecting data for such a need is extremely challenging. Taking that data and transforming it into something which decision-makers understand is the aspect that we will explore, specifically looking into how this was tackled on one of the SLS studies: SLS Asthma.

THE SLS ASTHMA STUDY

A total of 4,233 patients were recruited to the Asthma study which was set in Salford and Southern Manchester, UK. The study had just six inclusion criteria:

- Signed informed consent
- Have General Practice (GP) diagnosis of asthma
- Must be taking inhaled corticosteroid (ICS) monotherapy, or in combination with a long-acting beta2-agonist (LABA), for at least 4 weeks prior to the randomisation visit, had symptoms as defined in the protocol in the past week prior to the randomisation visit
- Must be able to complete an electronic questionnaire
- Aged 18 or over at screening (if female, they must also be of non-child bearing potential or have a negative test at the randomisation visit).

There were also only nine exclusion criteria (such as no history of hypersensitivity to the study drug or who were currently taking Relvar Ellipta) which were defined to protect the scientific integrity of the study, regulatory acceptability or patient safety.

The intention of the SLS Asthma study was to maintain the scientific rigor of a traditional RCT by utilising a randomised design and having a control arm. At the same time, the aim was to keep it as near to everyday clinical

practice as possible, that is, to have minimal inclusion/exclusion criteria, to keep the patient experience as normal as possible and to collect endpoints relevant to patients and healthcare decision makers.

This open-label study was designed to evaluate the effectiveness of initiating treatment with Relvar Ellipta (referred to as FF/VI for the remainder of this paper) compared to continuing treatment with 'Usual Care' (UC), where a physician was free to choose the appropriate asthma treatment for each patient based on their clinical judgement with licensed drugs. Patients would attend a randomisation clinic visit, consent to participate in the trial, and be randomised to initiate treatment with FF/VI or continue UC arm for a period of 12-months; in line with usual practice patients were permitted to modify treatment and remain in the study the only exception being patients randomised to UC were not permitted to take FF/VI and remain in the study. There was no protocol-mandated clinic visits during the 12-month treatment period, other than the screening and randomisation visit(s) at the start of the study and the Month 12 visit at the end of the study. Phone calls would be conducted at Weeks 12, 24, and 40 to complete the primary endpoint questionnaire – Asthma Control Test (ACT) – and to provide information of serious adverse events (SAEs) and adverse drug reactions (ADRs) they may have experienced. Patients would then go about their normal lives, visit their GP when needed, and pick up their medication from their pharmacy routinely, relaxing the methods of a RCT and keeping the setting as real as possible.

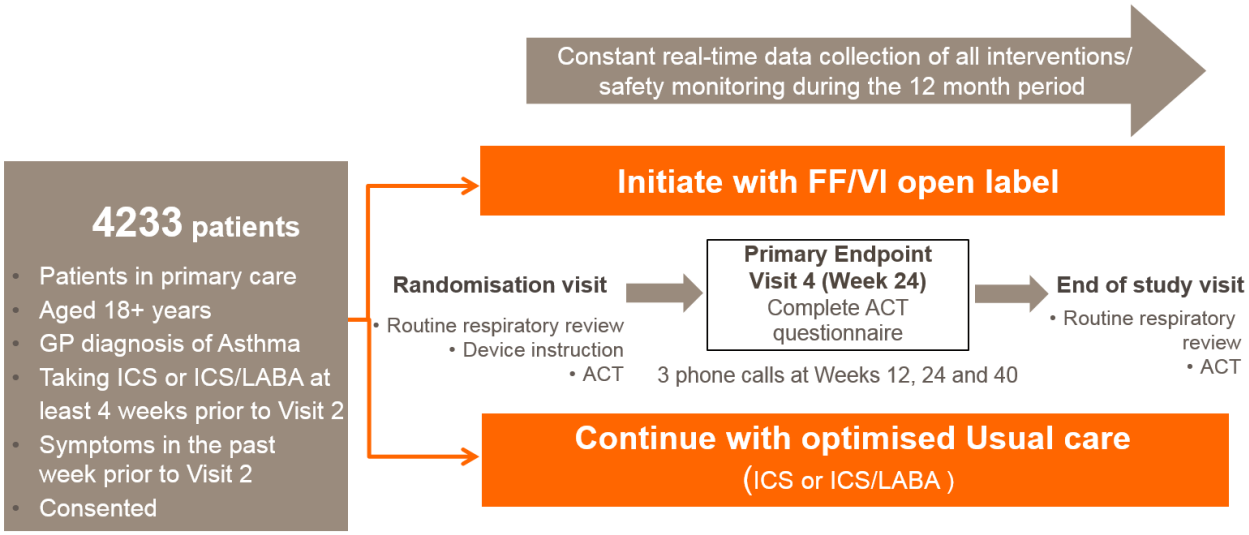

Figure 1: Study Design

HOW DATA WAS CAPTURED

Electronic Medical Records (EMR) data collected in this study are used to monitor safety and robustly collect data and these come from many systems and in various formats from across the UK National Health Service (NHS) and other healthcare systems. Each of these data feeds are owned by different organisations. Daily extracts are performed to give the most up-to-date information. The Linked Database System (LDS) developed by an external healthcare vendor combines this data and gives patients a unique subject ID as per the study in order to anonymise them by removing any personally identifiable information.

Baseline, demographic and other data such as SAEs and ADRs, study medication prescriptions were collected on the CRF. Questionnaire data was entered in the sitepads during the study.

Once the data was collated in this complicated process, it was then extracted as raw SAS datasets to GSK ready for statistics and programming to create the analysis datasets and Tables, Listings and Figures (TFLs).

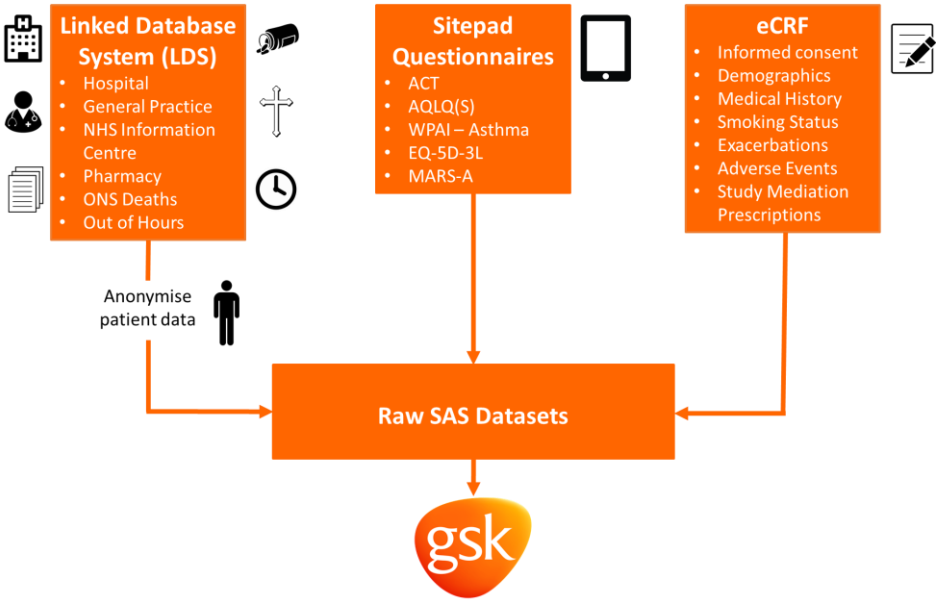

Figure 2: Data Capture
ONS = Office of National Statistics

CHALLENGES WITH THE DATA

This discussion centres around the challenges with using and analysing the raw datasets. Due to the complex nature of the data, there were a vast number of challenges to achieving the analysis that was required. Data which came from patients' EMR via LDS, such as primary or secondary care or prescriptions from pharmacies could not be queried or changed. This posed many challenges, such as:

- Potentially duplicate data – both true duplicates and false duplicates which can be indistinguishable
- Missing data – how do we approach this and how was is factored into the analysis
- Free-text fields – how do we consider these fields when it comes to programming
- Vast amounts of data – sieving through usually gigabytes worth of data for each dataset
- Filtering out the information which is required – considering which fields/columns/variables are of interest and targeting these to be analysed
- Data extracted into multiple datasets – unlike CRF data which is usually collected concisely for a clinical trial, EMR data is not, and could lie in different datasets, so merging potentially 'unclean' data is required
- Data from EMR is not collected primarily for clinical trial research – whereas the CRF is collected specifically for the clinical trial
- Lots of different coding systems and coding term levels – need to understand how each of the different coding systems work and how to interpret

Some examples of how we handled the raw datasets on the SLS Asthma study are described below to give an idea of the complex nature of the study. All the data in the examples in this paper are fictitious and are for illustrative purposes.

EXAMPLE 1: SCRAMBLING TREATMENT DATA

In a real-world study, and in the design of this trial, patients needed to know what they were taking and GPs and pharmacies needed to know what they were prescribing and dispensing. To maintain the blind for analysis and to avoid bias in the interpretation, analysis and development of the Reporting Analysis Plan (RAP or Statistical Analysis Plan – SAP) and TFLs, a form of data scrambling was required. An exposure dataset based on the CRF collected the study drug that the GP intended the patients to take. This would break the blind as it would identify the drug name, how much dose and how frequently they took it. FF/VI is a single inhaler which is administered as one puff once a day. For some patients in UC, multiple inhalers can be taken simultaneously, which results in multiple rows in the dataset. Even if the drug name, dose, frequency and number of puffs was masked, we would potentially still have a good idea of which treatment arm the patient was on.

One method that was implemented was to scramble the exposure dates and all study drug information in exposure across the patients using an independent unblinded programmer who was purely involved in this aspect of the study. The statistics and programming team would then remain blinded throughout the conduct of the study and GSK did not assess treatment differences until the final database was locked and the analyses was complete.

As an example, if a study had only three patients (see Figure 3), we would move the patient's exposure information onto the next patient below such that the exposure dates, drug name and dosing variables for subject 1001 drops to subject 1002, then subject 1002's exposure drops to subject 1003, and then subject 1003's exposure circles back to subject 1001 (although, in reality, an undisclosed random seed was used to maintain the blind). However, doing this would not sufficiently meet our needs.

Subject ID	Exposure Start Date	Exposure End Date	Drug Name	Dose	Frequency of dosing code	# of puffs per dose
1001	05-Jan-15	05-Jan-16	FF/VI	100/25	OD	1
1002	11-Feb-15	20-Sep-15	UC Drug A	100	BID	2
1002	11-Feb-15	20-Sep-15	UC Drug B	25	BID	2
1003	19-Oct-16	02-Nov-16	UC Drug A	50	BID	1
1003	03-Nov-16	03-Feb-17	UC Drug B	100	BID	4
1003	04-Feb-17	10-Feb-17	UC Drug B	100	BID	2
1003	11-Feb-17	20-Oct-17	UC Drug C	50	BID	1

Figure 3: Exposure Data Prior to Scrambling

As we can see in Figure 3, subject 1001 was exposed to the drug from 05-Jan-15 to 05-Jan-16. By scrambling subject 1003's exposure data to subject 1001, their exposure would be from 19-Oct-2016 to 20-Oct-2017 (Figure 4). We could consider this to be the end of the scrambling process, however, other data such as adverse events, would fall outside of this patient's treatment phase which would make the pre-programming more difficult to conduct if no adverse event was considered on-treatment for this patient and is temporarily being defined as pre- or post-treatment. Hence, the additional step we included was to keep the subject's start of exposure date.

Subj. ID	Subject's Date of First Exposure	Con-stant	Initial Scrambled Exposure Start Date	Initial Scrambled Exposure End Date	Final Scrambled Exposure Start Date	Final Scrambled Exposure End Date	Drug Name	Dose	Freq. of dosing	# of puffs per dose
1001	05-Jan-15	653	19-Oct-16	02-Nov-16	**05-Jan-15**	**19-Jan-15**	UC Drug A	50	BID	1
1001	05-Jan-15	653	03-Nov-16	03-Feb-17	**20-Jan-15**	**22-Apr-15**	UC Drug B	100	BID	4
1001	05-Jan-15	653	04-Feb-17	10-Feb-17	**23-Apr-15**	**29-Apr-15**	UC Drug B	100	BID	2
1001	05-Jan-15	653	11-Feb-17	20-Oct-17	**30-Apr-15**	**06-Jan-16**	UC Drug C	50	BID	1
1002	11-Feb-15	-37	05-Jan-15	05-Jan-16	**11-Feb-15**	**11-Feb-16**	FF/VI	100/25	OD	1
1003	19-Oct-16	-616	11-Feb-15	20-Sep-15	**19-Oct-16**	**28-May-17**	UC Drug A	100	BID	2
1003	19-Oct-16	-616	11-Feb-15	20-Sep-15	**19-Oct-16**	**28-May-17**	UC Drug B	25	BID	2

Figure 4: Exposure Data Post-Scrambling with Calculations

We then work out the difference between the subject's start date of exposure versus the scrambled subject's first date of exposure to get a "constant". We then use that "constant" to obtain the other exposure dates which are more relevant. For example, we have established that subject 1001's first date of exposure is 05-Jan-2015 and exposure data from 1003 has been scrambled to this subject. The difference between 05-Jan-2015 and 19-Oct-2016 (Subject 1003's first exposure date) is 653 days – the "constant". We then take the scrambled subject's end date of exposure, the next one being 02-Nov-2016, and subtract the "constant", to get 19-Jan-2015 – the scrambled exposure date for subject 1001. We then repeat the subtraction of the constant for the other exposure dates to get more relevant scrambled dates. Figure 4 describes the initial scrambled exposure dates against the final scrambled exposure dates. This process was repeated at each data transfer.

EXAMPLE 2: NUMBER OF PRIMARY CARE CONTACTS

One endpoint on the study was to compare how many primary care contacts a patient had between the randomised treatment arms. In addition, it was of interest to see which of those primary care contacts were asthma-related. This can provide an indication as to how much primary care resource a patient may consume from the NHS depending on the treatment arm a patient initiated on.

Primary care relates to GPs and other services which are often the first point of contact for healthcare advice or treatment. Data for this was extracted from different sources and feeds into two raw datasets – PC1HRU and PC2HRU.

The first – PC1HRU – contains data on when primary care contacts occurred and who they were seen by. Figure 5 is a simplified snapshot for an example patient.

Subj. ID	Date of Primary Care	Seen By
1001	03-Jul-15	System Administrator
1001	08-Jul-15	Receptionist
1001	19-Jul-15	General Medical Practitioner
1001	27-Jul-15	System Administrator
1001	30-Jul-15	Receptionist

Figure 5: PC1HRU Data

This list of seen by staff in Figure 5 were categorised into four groups following discussions with physicians and real-world evidence specialists: GPs, Nurses, Other and Admin. There were over a hundred different staff roles seen in the study, of which a small selection are shown in Figure 6. We consider any patients who were seen by GP, Nurse or Other as primary care contacts.

GP	Nurse	Other	Admin
Associate Practitioner	Associate Practitioner - Nurse	Acupuncturist	Administrator
Associate Practitioner - General Practitioner	Community Nurse	Chiropodist/Podiatrist	Appointments Clerk
GP Registrar	Community Psychiatric Nurse	Health Care Support Worker	Assistant
General Medical Practitioner	Contact Tracing Nurse	Health Professional Access Role	Audit Manager
House Officer - Pre Registration	Enrolled Nurse	Healthcare Assistant	Community Administrator
Locum	Midwife	Other Allied Health Professionals	Computer Manager
Partner	Nurse Consultant	Other Health Care Professional	Health Records Administrator
Senior Partner	Other Nursing & Midwifery	Other Students	Other Admin & Clerical
Sessional GP	Practice Nurse	Pharmacist	Receptionist
Sole Practitioner	Specialist Nurse Practitioner	Physiotherapist	System Administrator

Primary Care Contacts

Figure 6: Seen By Examples

The second dataset shown in Figure 7 – PC2HRU – contained the Read Codes and Terms which were entered for each primary care contact made by patients. Read codes are a standard clinical coding system used by GPs in the UK to code diagnoses, clinical signs, administrative items, tests and findings as well as many other items – similar to how adverse events are coded using standard dictionaries.

Subj. ID	Date of Primary Care	Read Term	Read Code
1001	03-Jul-15	Letter from consultant	9N36-1
1001	17-Jul-15	SMS text message sent to patient	9N3G
1001	19-Jul-15	Medication review with patient	8B3x
1001	19-Jul-15	Acute exacerbation of asthma	H333
1001	19-Jul-15	SMS text message sent to patient	9N3G
1001	26-Jul-15	Bowel cancer screening programme: faecal occult blood result	6866
1001	26-Jul-15	Bowel cancer screening programme: administration	90w
1001	26-Jul-15	No response to bowel cancer screening programme invitation	90w2
1001	30-Jul-15	Letter encounter to patient	9N35

Figure 7: PC2HRU Data

The aim was to merge these two datasets to get our desired outcome of the number of primary care contacts. The common variables among these two datasets are the Subject ID and Date of Primary Care. In doing so, we obtain the following:

Subj. ID	Date of Primary Care	Seen By	Read Term	Read Code
1001	03-Jul-15	System Administrator	Letter from consultant	9N36-1
1001	08-Jul-15	Receptionist		
1001	17-Jul-15		SMS text message sent to patient	9N3G
1001	19-Jul-15	General Medical Practitioner	Medication review with patient	8B3x
1001	19-Jul-15	General Medical Practitioner	**Acute exacerbation of asthma**	**H333**
1001	19-Jul-15	General Medical Practitioner	SMS text message sent to patient	9N3G
1001	26-Jul-15		Bowel cancer screening programme: faecal occult blood result	6866
1001	26-Jul-15		Bowel cancer screening programme: administration	90w
1001	26-Jul-15		No response to bowel cancer screening programme invitation	90w2
1001	27-Jul-15	System Administrator		
1001	30-Jul-15	Receptionist	Letter encounter to patient	9N35

Figure 8: Merge of PC1HRU and PC2HRU

After merging by Subject ID and Date of Primary Care, we could link the Read code and terms with whom the patient was seen by in date order. Some records existed in one dataset but not the other. In summary, for this particular example in Figure 8, we see that they were sent a letter on 03-Jul-15 by the System Administrator, they were seen by the GP, where they recorded an asthma exacerbation. A bowel cancer screening program was sent to the patient (no seen by information) but no response was recorded by the patient, and they were later sent another letter by a member of Administration on 30-Jul-15. Some of these events happened on the same day. A GP might see the patient and record each of the different symptoms or tasks that were carried out, in particular this was noticed for laboratory tests where tens of records would be entered.

In our analysis, it was reasonable to assume that we count primary care contacts by unique date, therefore although this patient has three records of being seen by the GP on 19-Jul-15, we would count this as one primary care contact.

We previously mentioned that another interest was the number of asthma-related primary care contacts. Read codes which were deemed to be relevant, as defined by medical and epidemiology experts prior to unblinding, to eliminate any bias were mapped to the merged dataset above. Highlighted in bold is one such event of an asthma-related primary care contact.

EXAMPLE 3: NUMBER OF SECONDARY CARE CONTACTS

Secondary care is contact with hospitals which can be planned (such as operations or referrals from GPs), or unplanned (such as emergencies). This is divided into three types of contacts:

- Inpatient (requiring the patient to be admitted into hospital),
- Accident & Emergency – A&E (for emergency unplanned admittance), and
- Outpatient (for those which do not require admittance into hospital).

Secondary care contacts are therefore counted as the summation of these three types of contacts. Each of these contact types are also of interest as well as finding out whether they are asthma-related contacts.

Secondary Care Contacts = Inpatient Contacts + A&E Contacts + Outpatient Contacts

To start with, let's explore inpatient data. This has been provided into two separate datasets, ADMHRU1 and ADMHRU2. The former contained spell information (defined as a period of time spent in hospital) for a patient including which hospital they attended and the admission and discharge dates, as well as various admission codes, discharge codes and diagnoses. The latter, contained the more detailed episode information (a particular healthcare service, such as surgery or a scan) within each spell for patients and the various primary and various secondary ICD-10 codes (up to 13 levels can be recorded from 1-13).

A patient can therefore have more than one episode within a spell.

Figure 9 and Figure 10 illustrate a simplified example of the two datasets.

Subj. ID	Spell ID	Hospital Name	Hospital Admission Date	Hospital Discharge Date	Primary Diagnosis
1001	1	HOSPITAL 'A'	01-Feb-16	02-Feb-16	S3250
1001	1	HOSPITAL 'B'	18-Mar-16	19-Mar-16	T438
1001	1	HOSPITAL 'C'	19-Mar-16	01-Apr-16	F388

Figure 9: ADMHRU1 (Inpatient Contacts) Data

Other data recorded in ADMHRU1 but not presented included: Admission Type, Discharge Method, Secondary Diagnosis, Primary Procedure, HRG Code

Subj. ID	Spell ID	Hospital Name	Episode	Main Diagnosis ICD10 Code	Secondary Diagnosis 1 ICD10 Code
1001	1	HOSPITAL 'A'	1	S3250	W014
1001	1	HOSPITAL 'A'	2	S3250	K590
1001	1	HOSPITAL 'B'	1	T438	T435
1001	1	HOSPITAL 'C'	1	F388	-

Figure 10: ADMHRU2 (Inpatient Contacts) Data

Other data recorded in ADMHRU2 but not presented included: Secondary Diagnosis ICD10 Code 1-13

We need to merge these two datasets together in order to get the most sense out of it. This was done by using the common variables – Subject ID, Spell ID and Hospital Name – see Figure 11.

Subj. ID	Spell ID	Hospital Name	Episode	Hospital Admission Date	Hospital Discharge Date	Primary Diagnosis	Main Diagnosis ICD10 Code	Secondary Diagnosis 1 ICD10 Code
1001	1	HOSPITAL 'A'	1	01-Feb-16	02-Feb-16	S3250	S3250	W014
1001	1	HOSPITAL 'A'	2	01-Feb-16	02-Feb-16	S3250	S3250	K590
1001	1	HOSPITAL 'B'	1	18-Mar-16	19-Mar-16	T438	T438	T435
1001	1	HOSPITAL 'C'	1	19-Mar-16	01-Apr-16	F388	F388	-

Figure 11: Merge of ADMHRU1 and ADMHRU2

To establish the number of inpatient contacts, we took the admission and discharge dates to see whether there was any overlap. It was reasonable to assume that if a patient was transferred from one hospital to another this was counted as one contact. In this instance, the patient therefore had two inpatient contacts. ICD-10 codes refer to the 10th revision of the International Statistical Classification of Diseases and Related Health Problems and contains thousands of different codes for diseases, symptoms and abnormalities. As per the asthma-related primary care contacts a similar process of identifying asthma symptoms was established for the ICD-10 codes prior to unblinding (to avoid bias) to flag those contacts which were considered to be asthma-related inpatient contacts.

For our next contact type, A&E, a contact is defined as unique attendances to the hospital. All A&E entries were provided in the AEHRU dataset (see example in Figure 12), which recorded the hospital they attended, their arrival and departure date, and the Diagnosis National Code amongst other variables. The Diagnosis National Codes[1] used here are a list of broad conditions which are coded for A&E attendances. Our interest in this stems on whether the reason they attended A&E was due to asthma or not. In the example below, the patient has one A&E contact and the Diagnosis National Code 1 = 251 decodes to 'Bronchial Asthma', which we classified as asthma-related.

Subject ID	Spell ID	Hospital Name	Arrival Date	Departure Date	Diagnosis National Code 1
1001	1	HOSPITAL 'A'	27-Oct-15	27-Oct-15	**251**

Figure 12: AEHRU (A&E) Data

Other data recorded in ADMHRU2 but not presented included: Arrival Mode, Diagnosis National Code 2-4, Diagnosis Local Code 1-4, HRG Code, Location Name, Attendance Category Description, Attendance Disposal Description

Lastly, for outpatient, this is given in the OUTPHRU dataset. As previously mentioned, these are contacts which do not require admission into hospital, so whether a patient attended or did not attend was recorded. These attendances were classified using the Main Specialty Coding[2], which divides clinical work into unique classifications. In Figure 13 example below, we have patient 1001, who did not attend their outpatient visit, therefore this does not get classified as a contact. For patient 1002, they attended both visits. Using the specialty coding, we have identified that the visit in July was for 'Respiratory Medicine', which we additionally defined as being asthma-related. In total, subject 1001 has no outpatient contacts, and subject 1002 has two.

Subject ID	Attendance Date	Attendance Time	Attendance Indicator	Specialty National Code	Specialty National Code Description
1001	19-Feb-14	15:45	Did not attend	100	General Surgery
1002	03-Jul-14	13:00	**Attend**	**340**	**Respiratory Medicine**
1002	12-Oct-14	09:00	Attend	330	Dermatology

Figure 13: OUTPHRU (Outpatient) Data

Other data recorded in OUTPHRU but not presented included: Appointment Type, Consultant Specialty code, Contract serial code, Clinical Codes, New or Follow-up, Primary Diagnosis, Primary Procedure, Sub-specialty code, Transport, HRG Code, Medical Staff type

The mapping in Figure 14 illustrates how the different secondary care contact endpoints were derived from raw data and pulled together to define our endpoint. For each of the contact types, an intermediate dataset was created to merge and transform the raw EMR data which could then be analysed. After this, the derivation of secondary care contacts was then performed in a separate dataset on a parameter-level, such that there was one record per patient per endpoint derivation. It is worth noting that we were mainly interested in those contacts which were 'on-treatment'. For each of these contacts these were classified based on the admission, arrival, or attendance date.

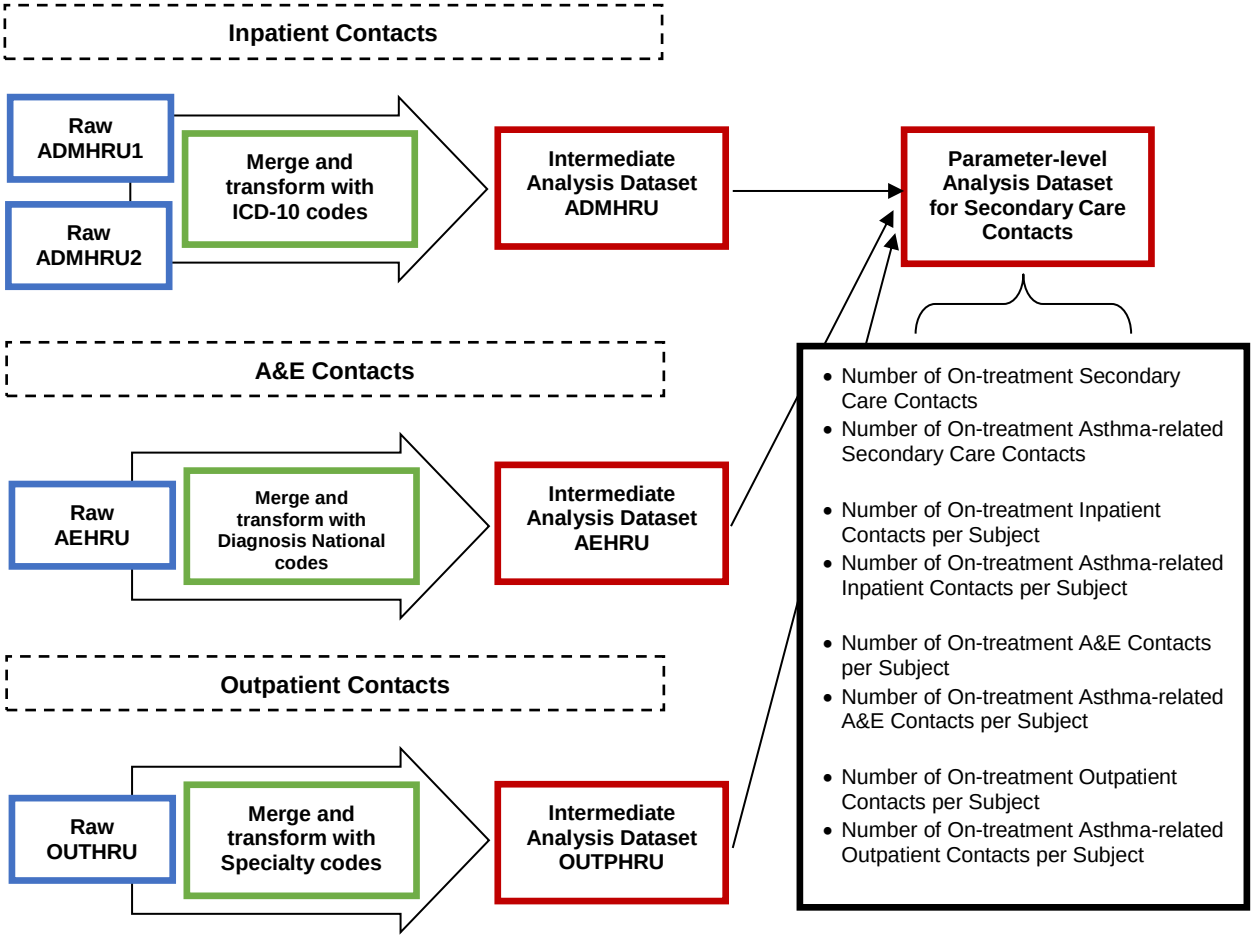

Figure 14: Mapping of Secondary Care Contacts

EXAMPLE 4: TREATMENT MODIFICATION

In the SLS Asthma study, to keep close to real-world scenarios, patients in either arm were allowed to modify treatment to any other licensed medication for asthma at any time during the study (although, as per protocol, patients who initiated on UC, were not allowed to take FF/VI) or modify the number of inhalations of the drug per day or the number of times (puffs) taken per day as advised by the GP.

Asthma inhalers are grouped into different 'classes' of medication. In this study, patients had to take a form of ICS, whether it be on its own or in combination (as separate inhalers or a combination inhaler) with LABA, which is

classified as an ICS/LABA. FF/VI is a form of ICS/LABA but for this study it is set to be in its own class. There are different brands within each class with different doses, dosing frequencies and number of puffs to be taken at each inhalation available. For example, an ICS inhaler might be administered as 100 mcg as one puff once a day, or as a 500 mcg dose per day administered as two puffs twice a day, depending on the licensing and recommendation of the GP.

To determine treatment modifications, we identified hierarchical steps to classify the type of modification (see Figure 15).

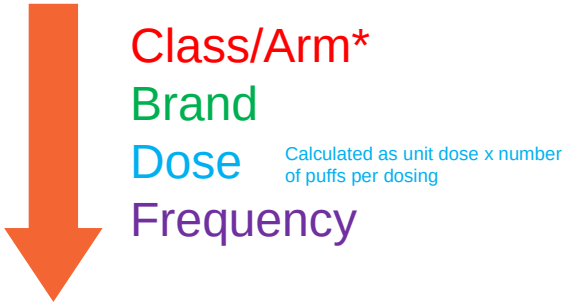

Figure 15: Hierarchy for Identifying Treatment Modification

Firstly, we assessed which class of medication they took and whether they stepped up in class (went from ICS to ICS/LABA), stepped down in class (from ICS/LABA to ICS) or switched treatment (from FF/VI to ICS or ICS/LABA but remained in FF/VI treatment group for the purposes of effectiveness analysis – or vice versa). If the patient did not change class, then we assessed within the class whether they switched brand of inhaler. If they were still on the same brand, we looked at whether they took a higher or lower dose. Finally, we looked at the frequency and whether this was a step up or down by calculating the daily number of puffs. If none of these criteria was fulfilled then there was no change to their medication. Figure 16 illustrates this in more detail.

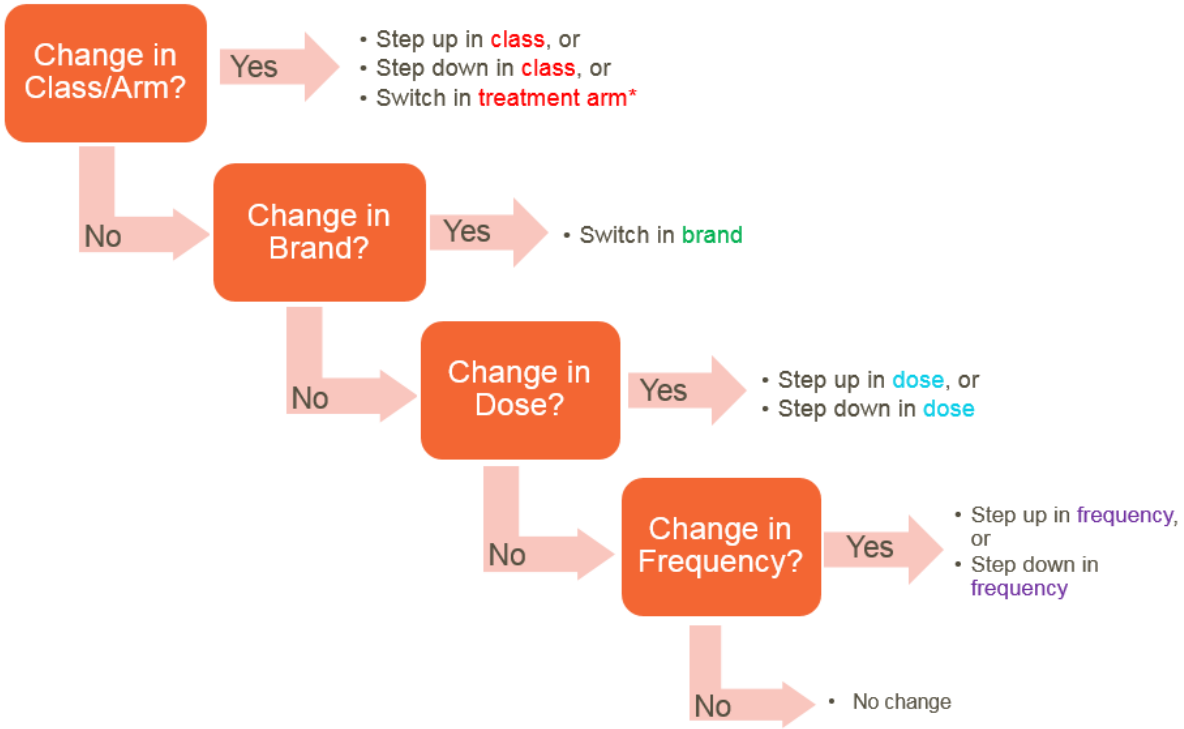

Figure 16: Decision Tree for Identifying Treatment Modifications

*Modify from FF/VI to UC treatment but remained in FF/VI treatment group for the purposes of effectiveness analysis

Data for exposure came from the CRF, which was recorded by the investigator to denote what the patient was prescribed to take in the study.

Consider the example below (Figure 17). In the CRF, the drug name, dose strength, frequency and number of puffs was provided as well as the dates of exposure, all in separate variables. For the purposes of describing the example, this has been simplified for ease of illustration. If multiple inhalers were taken during the same exposure date as another inhaler – i.e. took an ICS and LABA at the same time – then this was recorded in separate rows.

Sequence	Drug	Class	Modification
1	FF/VI 200/25 mcg OD (1 puff)	FF/VI	-
2	UC Brand 'C' 100 mcg BID (1 puff)	ICS	Switch in arm*
3	UC Brand 'F' 100/6 mcg BID (1 puff)	ICS/LABA	Step up in class
4	UC Brand 'S' 100/50 mcg BID (1 puff)	ICS/LABA	Switch in brand
5	UC Brand 'S' 100/50 mcg OD (1 puff)	ICS/LABA	Step down in frequency

Figure 17: Treatment Modification Example 1

A patient who was randomised to the FF/VI arm, was given a prescription for a UC Brand 'C' (ICS) and was classified as a switch in arm (although in the effectiveness analysis this patient would still be analysed according to the treatment they initiated on, i.e. FF/VI). The next inhaler they took was UC Brand 'F' (ICS/LABA) and since this is a combination drug with the addition of LABA, it is a step up in class. The patient then took a different ICS/LABA combination, UC Brand 'S', which therefore makes this a switch in brand. In their last modification there was a change in the frequency of UC Brand 'S' 100/50 mcg from one puff twice a day to once a day, making this a step down in frequency. In summary for this example, the patient had four modifications during the course of the study as described in the final Modification column with their treatment path (described by class) as FF/VI → ICS → ICS/LABA.

Sequence	Drug	Class	Modification
1	UC Brand 'Q' 100 mcg BID (2 puff) + UC Brand 'T' 25 mcg BID (2 puff)	ICS/LABA	-
2	UC Brand 'Y' 200/6 mcg BID (2 puff)	ICS/LABA	Switch in brand
3	UC Brand 'Y' 400/12 mcg BID (1 puff)	ICS/LABA	No change
4	UC Brand 'Y' 400/12 mcg BID (2 puff)	ICS/LABA	Step up in dose

Figure 18: Treatment Modification Example 2

In this second example (Figure 18) the patient took a form of ICS/LABA throughout. Notice that the first drug are two separate inhalers, an ICS and a LABA, though as previously described, this is still classed as ICS/LABA. The next drug taken was a combination ICS/LABA inhaler of UC Brand 'Y', and since this is no change in the class it therefore makes this a switch in brand. This inhaler has a dose strength of 200/6 mcg and was administered as two puffs. In the next change, the patient still took UC Brand 'Y', but at a dose strength of 400/12 mcg and was taken as one puff. Although this is administered one puff less, the dose strength is double the previous, therefore, taking this into account there was no change and no modification in this instance. In the last change, it appears that one puff of UC Brand 'Y' 400/12 mcg was not enough to control their asthma symptoms, so an extra puff of the same drug and dose strength makes this a step up in dose. This patient therefore has two modifications in total and a treatment path of ICS/LABA throughout.

CONCLUSION

The increasing trend and interest towards having more trials – or a need for more trials – in a real-world setting can yield results which can inform healthcare professionals in their decision-making processes. Being a ground-breaking study and a 'first of its kind' meant that the SLS study team faced a lot of difficult challenges. Only a few of the programming examples have been discussed here, but if we start to delve into how to code each of these examples from a programming point of view, we begin to get an idea of the sizeable task that we faced in using real world data.

Across the industry we use CRFs which are designed to capture data required for the analysis of clinical trials in a format that programmers and statisticians are familiar with. However, the EMR data was not captured in this way, it contained hundreds of variables which were not necessarily user friendly and could not be queried in the way we would normally expect.

Working with large datasets often with missing data added further complexities and challenges. To be able to understand the data that was supplied required close collaboration with statisticians, clinicians, physicians and epidemiologists to resolve issues prior to unblinding and before the database was locked. We needed to use care and consideration with the assumptions which were being made (e.g. how the different data sources would be merged together) to protect the integrity of the study.

It is imperative that programmers are on-board at the very beginning of these types of studies, to understand how we can take the raw data and transform it into something that can be used to answer the questions posed in the protocol.

When conducting a study of this nature, with complex and unfamiliar external data sources, it is important for programmers to understand the format of the data and potential issues in using and analysing it. Instream data or test data would aid and increase the awareness and understanding of the data and raise potential issues and therefore make more informed decisions at an earlier stage.

The unique nature of these studies presented a lot of challenges that the study team had not encountered before, it was nevertheless an exciting and rewarding experience to work in innovative ways on this pioneering project.

REFERENCES

[1] Accident & Emergency Diagnosis Codes:
https://www.datadictionary.nhs.uk/web_site_content/supporting_information/clinical_coding/accident_and_emergency_diagnosis_tables.asp

[2] Specialty Codes used for Outpatient Contacts:
https://www.datadictionary.nhs.uk/data_dictionary/attributes/m/main_specialty_code_de.asp

RECOMMENDED READING

Please see the following manuscript if you are interested in reading more about the study and the results:

Woodcock A, Vestbo J, Bakerly ND, New J, Gibson JM, McCorkindale S, Jones R, et al. Effectiveness of fluticasone furoate plus vilanterol on asthma control in clinical practice: an open-label, parallel group, randomised controlled trial. Lancet 2017;390:2247–2255

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Daniel Kwok
GlaxoSmithKline
Stock Park West, 1-3 Ironbridge Road,
Uxbridge, UB11 1BT
Email: daniel.k.kwok@gsk.com

Jennifer Colthart
GlaxoSmithKline
Stock Park West, 1-3 Ironbridge Road,
Uxbridge, UB11 1BT
Email: jenny.w.colthart@gsk.com