

PhUSE 2013

Paper HE01

My Big Data experience – getting my feet wet in the puddle and still treading water 2 years later

Rob Walls, F. Hoffman-La Roche Ltd., Basel, Switzerland

ABSTRACT

"Big Data" - Electronic Health Records(EHR) and Health Insurance Claims data are growing exponentially every day. Regulatory focus has widened to ensure approved pharmaceutical products are efficacious and safe in broader populations in the real world. This is an area of extreme interest in today's ever changing and increasingly competitive pharmaceutical industry but what exactly is it? Where does it come from? What does it contain? How can we use this type of data to inform on the safety of marketed products and during the development of pharmaceuticals?

My first two years of exposure since clinical trial programming have meant many perception shifts. I will explain data structures, touch on some SAS® tips and tricks, and describe the differences between EHR and Claims data. Most importantly I will show how the right analysis can help effectively inform and affect decision making across all stages of the molecule's development lifecycle and marketing.

INTRODUCTION

Access to Electronic Health Records (EHR) and Health Insurance Claims data is revolutionizing the Healthcare and Pharmaceutical industry. Though it might be a misnomer, these types of data are commonly being referred to as 'Big Data' but I prefer the term Real World Data (RWD).

I came from a background of programming clinical trials data using SAS® and two years ago I moved from this highly specialized area into working on RWD, mainly for Epidemiological research. The shift from one to the other has been far from the straight forward step that I thought that it would be.

With that in mind, this paper is aimed at data programmers currently working in the Clinical Trials area who would like to understand more about how this RWD is currently being used within pharmaceutical research. I am going to make broad comparisons between clinical trials data and studies and those of RWD. I will go on to talk about some of the exciting areas in which RWD can be put to use, helping to inform and contextualize the entire development lifecycle and strategy of our drug development processes, from early development phase right through to the post-marketing of our more mature products. Finally, a small amount of time will go into looking at some of the shifts in perception that I had to adopt in order to fully get to grips with working with this data (both programmatic and functional).

CLINICAL TRIALS VS. REAL WORLD DATA

CLINICAL TRIALS

In clinical trials, data collection is a carefully regulated and structured affair. Electronic Case Report Forms (e-CRF's) are designed and implemented in database design to ensure that data quality and consistency are maximized, allowing the scientists involved to be able to answer very specific, targeted questions. Standards have been put in place and are constantly evolving to ensure that this level of data quality is now an industry standard (CDISC) so that regardless of the company conducting the trial, one should know when looking at the data what each field represents. In a nutshell the clinical trials data is primary, fit for purpose data (which rarely gets above 10,000,000 observations) and is, to all intents and purposes, CLEAN.

PhUSE 2013

REAL WORLD DATA

Compare this now to RWD. The data structures are designed by the vendors for their own specific purposes. In Claims data this is to allow the vendor to keep track of medical insurance claims for reimbursement; in EHR's, the purpose is to capture all information on a patient's health and treatment history in order to have as much relevant information as possible to help the physician with decisions relating to the treatment of a patient. This means that RWD is secondary use data, and as such can contain inconsistencies and unexpected nuances that are not always immediately evident. The data structures are as varied as the vendors who supply them, and though they may sometimes use the same coding systems, the data itself can be laid out in vastly different structures. We have to approach this data in a careful and considered manner as some of the datasets can run into hundreds of Gigabytes (a single data set can be >7 billion observations).

ELECTRONIC HEALTH RECORDS (EHR) DATA

In this database type we have databases whose primary purpose is to support in the provision of healthcare to patients. Below are two of the more commonly used data sources available:

- Clinical Practice Research Datalink [CPRD®] – (>50m patient lives).
- General Electric - Electronic Medical Records [GE-EMR®] – (>25m patient lives).

Pros of using this type of data:

- Generally contain labs and lifestyle factors (BMI, Smoking status, etc.).
- Contain written prescriptions or drugs given during office visits, which is extensive.
- Potentially cover patients from birth to death.
- True denominator can be derived (by only considering 'active' patients).

Cons of using this type of data:

- Generally we have no payment information.
- Contains only prescriptions written (which is not the same as 'taken').
- Suitability of a data source can differ based on the data provider type.

Though both of these data sources contain a patient's medical records, the CPRD database is an example of a UK General Practitioners database where information comes from the primary healthcare provider (or GP). From this we have a very comprehensive overview of a patient's day to day medical life. Where this data source is limited is when we have to look into specialist treatment, for example Oncology, where the patient will be treated in a specialist clinic. It is very difficult to be able to 'retrieve' this sort of information from the traditional database, so what we are now seeing is that CPRD are linking their data with a whole host of external UK databases, including the Office of National Statistics (registries of births and deaths, socio-economic factors, etc.), Hospital Treatment Insights (HTI) and Hospital Episode Statistics (HES for hospital diagnoses).

An alternative type of EHR is one from a US healthcare network (for example, GE-EMR) where a series of 'networked' primary care provider data is collected and stored by a central provider who then provides this HIPAA compliant, anonymized data to external customers. This data is nationwide and on-the-hole representative of the US population however, because of the nature of the US insurance based healthcare system, patients can come and go between one provider and another for treatment and/or consultations. While the first provider may subscribe to this network the second may not. This means that it is not possible to get a true denominator from this type of database, limiting us to point estimates.

MEDICAL INSURANCE (CLAIMS) DATA

The primary purpose for claims data is for the payment of what a patient has claimed against their medical insurance. Because of this, the detail of information about a patient's treatment and care is very robust, crossing the GP/hospital/healthcare network boundaries which cause limitations to the EHR database types.

- Optum Insight InVision Data Mart® [US] – (>32m patient lives)
- Truven MarketScan® [US] – (>140m patient lives)

The above two data sources are very similar. There is a distinct start and stop date to each patient's periods of enrolment in the health insurance plan (including what type of coverage they have e.g for drugs, mental health, etc.) which can be used to help ensure that you are selecting viable patients for your study. For example, if running a Schizophrenia study, you would not include patients who did not have mental health coverage. These data sources also contain all diagnoses, procedures and medications (filled/dispensed prescriptions) from visits for all types of service as well as detailed payment information about medications and procedures. It should be noted however that

PhUSE 2013

when pricing/cost data exists it is normally derived from an algorithm in order to prevent identification of pricing strategies.

GENERAL CONSIDERATIONS

The main problem with secondary use data is gaining a clear understanding of what each variable actually contains. A number of examples of the intricacies that need to be considered are presented below:

- If a healthcare provider enters a diagnosis code relating to "Rheumatoid Arthritis", is this a straight diagnosis of RA, or is this perhaps the doctor ordering a set of screening labs to be run to rule out a diagnosis of RA? Perhaps it is better to consider only diagnoses that have a second 'confirmatory' diagnosis within a set timeframe after, or you could look at procedures conducted at the same visit. If there are blood draws or some other diagnostic procedure (e.g. MRI, X-ray, etc.) then perhaps this is an investigative diagnostic code and not a diagnosis proper. Maybe then we can rule out this instance, or do a combination of both approaches.
- In MarketScan, when a patient reaches the age of 90, in order for the vendor to remain HIPAA compliant they must never show patients in the data as being over that age. The vendor's solution is to increment the patients' date of birth for every new year after reaching 90 that they are in the data. Our problem is that they don't tell you this. It is only when starting to work with the data that you notice these patient inconsistencies and start asking questions.
- Death dates: In CPRD some patients are marked as dead who later have several years' worth of data showing that the patient is indeed still alive, due to the fact that data is entered by GPs and is not always correct. In MarketScan it is possible to find a death date for some patients but this cannot be relied upon either as these are captured only when the death occurs in certain clinics or hospitals and not when a patient dies outside these institutions. How do you then come up with a denominator for this?

Finally, we, as programmers, need to consider how best to store and access the data. Due to its size, we need to balance up ease of use with robustness so do we try to bring all of this data together under one over-arching model so that standard programming can be used for extraction, meaning an overhead in initial processing time and a potential loss of sensitivity in the data? Or do we respect the differences of each data source and develop programs to run directly on each database?

CLINICAL TRIALS VS. REAL WORLD STUDIES

When we conduct clinical trials, our ultimate goal is hypothesis testing. The studies are designed to answer set scientific questions through the recruitment of clinically homogenous patients. Studies are planned and laid out according to strict protocols to provide the evidence behind the science. Companies have full control over each and every data point and how it is generated and there are clearly defined timelines around each data delivery.

With RWD we are involved in hypothesis generation. Studies can be conducted in order to fulfill a varied number of purposes which will support drug development, sometimes from the definition of the original therapeutic area all the way through to marketing strategies

Firstly when designing a study you need to carefully consider which database(s) may be suitable. There is no point using CPRD for an oncology study for example, as this is primary caregiver data and will not contain much in the way of cancer diagnoses or treatments.

Since each study has its own unique question and each data source has its own idiosyncrasies, a decision has to be made on how best to proceed. The pros and the cons of using database A against that of database B need to be weighed up. Which database will give you a more robust, 'realistic' answer to your question? Generally, there is no right or wrong answer on how to handle or approach each specific question. All we can do is ensure consideration is given to the choice of data source and data extrapolation in order to answer these questions as best we can.

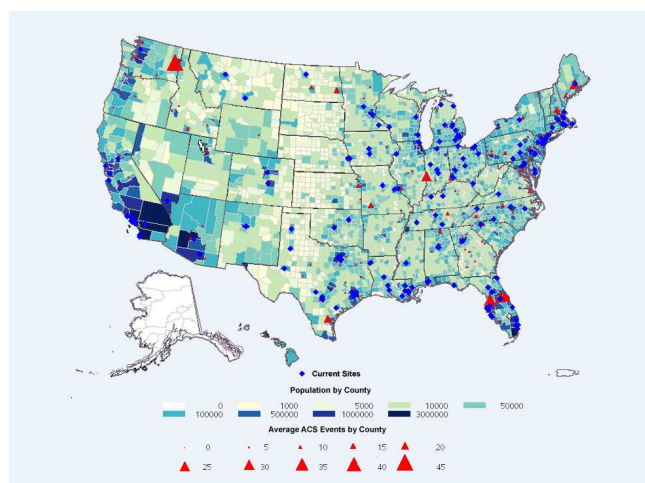
RWD has any number of uses in the clinical drug development lifecycle. I will now go on to talk about some of the more widespread and current uses of this type of data.

PhUSE 2013

CLINICAL TRIAL PATIENT RECRUITMENT

POTENTIAL RECRUITMENT OPPORTUNITIES

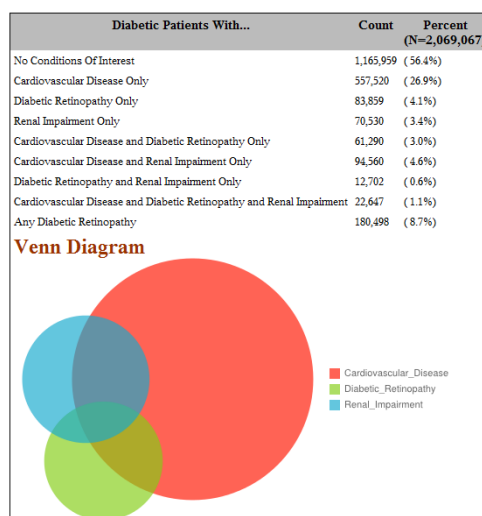
Using RWD we can aid in the identification of likely study populations. Heat maps can be generated indicating where disease hotspots can be found. Pictured below is a map of the US with each county shaded according to population density. The blue diamonds represent current recruiting sites while the red triangles show the average Acute Coronary Syndrome (ACS) events for any particular county. This allows the project team to be able to more readily target sites in hotspot areas that are currently not being taken advantage of.



RECRUITMENT OVERLAPS

We can also identify potential overlaps in clinical trial recruitment. So, for example, if two studies are recruiting in parallel for the same indication but for slightly different subpopulations then there is the potential for conflict where they might both be vying for recruitment of the same patients. Investigation into RWD sources can help in the initial stages by providing a guide as to what scale of overlap can be expected, allowing the two trial teams to understand whether they need to target different sites, or whether any overlap in the patient populations is negligible and can be ignored.

Presented below is an overlap analysis for diabetic patients with cardiac disease, renal impairment and diabetic retinopathy.



PhUSE 2013

SAFETY INPUT

SAFETY SURVEILLANCE

As pharmaceutical companies are scrutinized more and more closely we are rightly more accountable for the maintenance of the safety profiles of our drugs. One way to do this is to proactively monitor our compounds once they reach the market. Once they are on the market, safety data will invariably be recorded in one RWD source or another. This allows us to analyze drugs as they are prescribed and used, beyond the 'sterile' clinical environment. Analyses of this sort can lead to drugs being withdrawn from market (in the most severe instance) long before it might normally have happened, thereby minimizing patient risk, or lead to label updates, which could be either positive or negative.

CLINICAL TRIAL SUBMISSIONS

Keeping a careful eye on safety signals that are surfacing within patients treated with our drugs is key in clinical trials, however, it is not always known what to expect from a population of patients with indication X being treated with drug Y. By mining into RWD we can look into non-drug Y exposed patients and create a profile of safety for this indication, outlining what sort of comorbidities can be expected. If there is an increased risk of coronary heart failure in real world data populations, then when you see these signals surfacing in the clinical trials then this evidence can put this signal into context when submitting to the regulators. We can state that this signal is a comorbidity of the disease and therefor unlikely to be a side-effect of the treatment.

INVESTIGATIONS OF OFF-LABEL TREATMENT PATTERNS

There is now a new requirement from the EMA for pharmaceutical companies to provide details of 'off-label' drug usage. This is an obvious application of RWD but is not as straight forward as it sounds. The problems arise when you try to define 'off-label' usage.

Treating physicians have set drugs which they know are covered by their payer for treating certain diseases. When the doctor decides that a patient may also, in their clinical opinion, benefit from the off-label use of a drug then they may 'diagnose' the patient as having the suitable disease beforehand so that the patients records show that they are eligible for the treatment and hence the costs will be reimbursed. How do we identify this? A careful algorithm has to be put in place to take payment and prescribing patterns into account as well as considerations for the disease itself. Secondly, once you have identified 'off-label' use then it is necessary to work backwards through the patients' history and somehow divine from the data what this prescription has been given for.

HEALTH ECONOMICS

When going to funding bodies, such as NICE, it is important to be able to justify our claims as to why a drug should be added to their formularies. It is all very well to have a drug whose safety and efficacy has been clinically proven, but if it is seen as being overly expensive then an argument is needed as to why they should pay the extra, for what may be seen as only a marginal benefit to a small group of patients. Since the drug is not going to be available in that country yet then there are a couple of approaches that can be taken. We can run cost/benefit ratios of the current market using data for the current standard of care and then forecast how the drug would fit into this based on clinical evidence. We can also, if possible, run cost/benefit analysis in data from other countries where the drug may already have been approved for payment and assuming that costs and treatment methods are comparable, make an argument from this.

It gets even more complicated when considering this impact going across therapeutic areas. Unfortunately, most payers do not have limitless funds and an increase in cost here has to mean a decrease somewhere else. Therefore it is no longer enough to say to payers that, though it costs more, diabetes drug X is more efficient and safer than diabetes drug Y so therefore should be the standard of care. The increased cost of drug X against the impact this increase in expenditure will have in other therapeutic areas needs to also be weighed up. You might actually end up vying for funding against one of your own oncology drugs as it is now also a question of, for example, does the risk/cost/benefit to diabetes patients being treated with drug X outweigh the risk/cost/benefit of drug Z for the treatment Non-Hodgkin's Lymphoma?

Unfortunately, due to the data we are generally (but not always) limited to the use of claims databases in order to do this since they have the richest and most complete payment and charge information.

PhUSE 2013

MOLECULE DEVELOPMENT

There is no denying that clinical trials are a long, drawn out and costly process, therefore it is important when making decisions on the molecule level, to have a understanding of the cost of trials versus the potential risk/benefit for patients. We can inform decisions to end development of a molecule early thereby saving money, or conversely, argue for the continuance of development where a unique patient value is seen in RWD. This can very successfully be done by profiling the disease areas before the full clinical trials get underway so that the development teams have a clearly defined picture of the disease area. This will allow them to make better, more informed decisions because they know more about the intended patient populations.

RISK ESTIMATIONS

With every chemical or biological compound that we put into our bodies, be it food, drink or medicinal there comes an associated risk. This risk associated with pharmaceuticals and biologics has long been recognized as requiring careful monitoring and regulation. Therefore every year more and more is expected of pharmaceutical companies in order to justify the risks of their compounds and thereby enabling them to get to or stay on the market.

Each molecule that comes to market now requires a Risk Management Plan (RMP) to be developed and once on the market a yearly Periodic Benefit-Risk Evaluation Report (PBRER) has to be submitted. For this, it is necessary to monitor the risks associated within the 'at risk' populations of patients taking the drug, or with certain diseases in order to gauge the potential impacts that our drugs are having on a population. For example, if patients in a disease population have a higher risk of cardiovascular disease (CVD), then is it ethical to give them a drug that we know might increase their chances of stroke? Does the potential benefit to a patient outweigh the risks involved? If the disease being treated is, say eczema then perhaps not, but if it is a potentially life-threatening disease then perhaps the small increase in risk of CVD is preferable.

COMPARITIVE EFFECTIVENESS

Comparative effectiveness research does exactly what it says. It compares the effectiveness of drug A against a competing drug B, and because of the wealth of RWD to which we have access, we can look into the overall efficacy, the costs involved in treatment, the risk/benefit profiles of both drugs as well as regional variations of prescription/payer approval. We can see how our compounds operate in the real world and actually build up a picture of gaps in marketing strategy, why our compounds are perhaps not being prescribed in certain regions or have some payer restrictions associated. This allows us to actively target the placement of our drugs better and more effectively on the market to better benefit our patients.

SHIFTS IN PERCEPTION

From a personal perspective, one of my preconceptions that I had to reconsider is that CLEAN data equals GOOD data. EHR and Claims data, can be unclear. There are gaps in information, inconsistent ages and genders for patients as well as the aforementioned nuances and inconsistencies, all of which have to be carefully considered before starting on any analyses. During my time working on clinical trials these types of irregularity would be deemed as completely unacceptable. The difference in tolerance comes from the fact that we are dealing with much larger populations and the larger the population the less impact on results due to a few missing or incomplete variable values. Since our aim is hypothesis generation then this can be acceptable.

Programming with clinical trials data is linear and generally you will follow a molecule through the various phases of its development, so you get to know the science and studies of your compound in great detail. However, because of this we often forget the full scope of what is going on outside of clinical trials. A much greater knowledge of the development lifecycle and post-approval activities is required when working with RWD as we have to handle requests that could come in from a multitude of functions within the company, supporting various demands, both internally, academically and also for regulators. I have lost the fine detail that I once had but have now been exposed to a much broader and, I think, more challenging string of questions to answer. We need to know how the therapeutic areas overlap and interact, where each molecule is in its development so that we can prepare for what type of questions may come in, while also keeping an eye on post-marketing support of more mature products.

PhUSE 2013

SAS EFFICIENCY CONSIDERATIONS

I always thought that I was an efficient coder but when I started to work with RWD I quickly became aware of the gaps in my knowledge. Here it is not only efficiency of processing that is needed but, since every kilobyte counts, we have to be aware of streamlining data whenever possible as this can have profound impacts on run times. For example, formatting of single length character fields is more efficient than formatting a numeric field.

When coding using RWD we really have to adhere to strict coding efficiencies. Some general principals are listed below:

- Where possible, use only one pass of the data to extract everything that you need.
- Minimize the number of data/proc steps that you use.
- Consideration of the Logical Program Data Vector in order to best utilize in data step processing.
- Only read in the variables and observations that are absolutely necessary.
- Avoid SCAN, SUBSTR and INDEX functions where possible as these are processor intensive.
- Sort as infrequently as possible.
- When merging data, check which technique is most efficient out of MERGE, SQL and SET KEY. In my experience, with the volumes of data we use, the traditional SAS MERGE generally comes out on top.
- Maintain your variable lengths to be as small as possible.
- Maintain your variable types to minimize storage (and hence processing), for example a formatted single length character field is much more efficient than a formatted numeric field.

Indexes are also a very useful tool to help us get the data we want very quickly, but there are also considerations here that should be taken into account. Due to the overhead in processing time required in their creation, it is only an option to create these against static data sets designed to be used multiple times. Another consideration is the ordering of variables in the data. This can have profound advantages on data access, and though it may sometimes require creating more than one version of a data set (with different ordering), the efficiencies may often outweigh the overhead in creation.

In the example below I have shown the log times for a traditional pull of all patient diagnoses of 'Warts' (ICD-9 code: 078.***) within the MarketScan Commercial and Medicare databases 1999-2011. In the traditional MarketScan structure, diagnoses are captured across multiple data sets and in multiple diagnosis variables per data set per visit for a patient. Obviously we can expect this pull not to be the most efficient.

```
*****
** Processing for: WARTS
** ICD Codes:      078##
** Proc Codes:
** ICD9-CM Procs:
** DRG Codes:
** Revenue Codes:
** CPT Codes:
** HCPCS:
*****,
```

```
NOTE: There were 801 observations read from the data set DMK_SCAN.CS_1999.
      WHERE (DX1=:'078') or (DX2=:'078');
```

```
NOTE: There were 110 observations read from the data set DMK_SCAN.RS_1999.
      WHERE (DX1=:'078') or (DX2=:'078');
```

```
NOTE: There were 220295 observations read from the data set DMK_SCAN.CO_1999.
      WHERE (DX1=:'078') or (DX2=:'078');
```

```
NOTE: There were 14001 observations read from the data set DMK_SCAN.RO_1999.
      WHERE (DX1=:'078') or (DX2=:'078');
```

.....

```
NOTE: There were 29014 observations read from the data set DMK_SCAN.CS_2011.
      WHERE (DX1=:'078') or (DX2=:'078') or (DX3=:'078') or (DX4=:'078');
```

```
NOTE: There were 4085 observations read from the data set DMK_SCAN.RS_2011.
      WHERE (DX1=:'078') or (DX2=:'078') or (DX3=:'078') or (DX4=:'078');
```

```
NOTE: There were 1873775 observations read from the data set DMK_SCAN.CO_2011.
      WHERE (DX1=:'078') or (DX2=:'078') or (DX3=:'078') or (DX4=:'078');
```

```
NOTE: There were 135301 observations read from the data set DMK_SCAN.RO_2011.
      WHERE (DX1=:'078') or (DX2=:'078') or (DX3=:'078') or (DX4=:'078');
```

PhUSE 2013

NOTE: The data set SOURCE._01A_WARTS_1999_2011_S0_COMMCA has 15265014 observations and 13 variables.

NOTE: DATA statement used (Total process time):

real time	5:31:38.78
cpu time	4:03:55.74

So a 5.5 hour run-time to extract a small-medium sized population. With a little bit of restructuring, we were able to bring all diagnosis variables into a single variable and we combined all yearly data into the same data set so that we only had one streamlined data set to deal with. We then sorted and indexed this on ENROLID (unique patient identifier) and CODE (the diagnosis codes in question). These were the results:

NOTE: There were 9119144 observations read from the data set DFAST_CA.DIAGS.

```
WHERE code in ('078', '0780', '07800', '07801', '07802', '07803', '07804',
'07805', '07806', '07807', '07808', '07809', '0781', '07810', '07811', '07812',
'07813', '07814', '07815', '07816', '07817', '07818', '07819', '0782', '07820',
'07821', '07822', '07823', '07824', '07825', '07826', '07827', '07828', '07829',
'0783', '07830', '07831', '07832', '07833', '07834', '07835', '07836', '07837',
'07838', '07839', '0784', '07840', '07841', '07842', '07843', '07844', '07845',
'07846', '07847', '07848', '07849', '0785', '07850', '07851', '07852', '07853',
'07854', '07855', '07856', '07857', '07858', '07859', '0786', '07860', '07861',
'07862', '07863', '07864', '07865', '07866', '07867', '07868', '07869', '0787',
'07870', '07871', '07872', '07873', '07874', '07875', '07876', '07877', '07878',
'07879', '0788', '07880', '07881', '07882', '07883', '07884', '07885', '07886',
'07887', '07888', '07889', '0789', '07890', '07891', '07892', '07893', '07894',
'07895', '07896', '07897', '07898', '07899');
```

NOTE: The data set SOURCE._02A_WARTS_1999_2011 has 9119144 observations and 4 variables.

NOTE: DATA statement used (Total process time):

real time	2:04:30.31
cpu time	18:29.56

By doing this we have reduced our run time by 62% from 5.5 to 2 hours; a significant improvement! However, we then decided to try the resorted approach and next ordered and indexed the data on CODE (the diagnosis codes in question) and then ENROLID (unique patient identifier), ran it through the same extract program to find:

NOTE: There were 9119144 observations read from the data set NEW_CA.DIAGS_CODE.

```
WHERE code in ('078', '0780', '07800', '07801', '07802', '07803', '07804',
'07805', '07806', '07807', '07808', '07809', '0781', '07810', '07811', '07812',
'07813', '07814', '07815', '07816', '07817', '07818', '07819', '0782', '07820',
'07821', '07822', '07823', '07824', '07825', '07826', '07827', '07828', '07829',
'0783', '07830', '07831', '07832', '07833', '07834', '07835', '07836', '07837',
'07838', '07839', '0784', '07840', '07841', '07842', '07843', '07844', '07845',
'07846', '07847', '07848', '07849', '0785', '07850', '07851', '07852', '07853',
'07854', '07855', '07856', '07857', '07858', '07859', '0786', '07860', '07861',
'07862', '07863', '07864', '07865', '07866', '07867', '07868', '07869', '0787',
'07870', '07871', '07872', '07873', '07874', '07875', '07876', '07877', '07878',
'07879', '0788', '07880', '07881', '07882', '07883', '07884', '07885', '07886',
'07887', '07888', '07889', '0789', '07890', '07891', '07892', '07893', '07894',
'07895', '07896', '07897', '07898', '07899');
```

NOTE: The data set SOURCE._03A_WARTS_1999_2011 has 9119144 observations and 4 variables.

NOTE: DATA statement used (Total process time):

real time	50.56 seconds
cpu time	1:12.62

This transformed the results which went from 5.5 hours to 50 seconds, an improvement of 99.75%, demonstrating how having a clear understanding of indexes and how they store the reference data is extremely important to be able to work efficiently.

There are other ways that SAS can be made to perform more efficiently, through the use of the HASH object, or adjustments to the number and size of buffers assigned, but these are not going to be covered.

PhUSE 2013

CONCLUSION

The potential impact of the use of RWD is staggering. We have access to millions of patients' lives and can data mine in order to answer a variety of questions; we just have to know the limitations of our data so that the question being answered is as transparent and impactful as possible. We can directly benefit patients by ensuring that drugs get to market quicker and ensure that treatments are more successfully targeted to suitable patients. We can also help with the strategic placement of drugs on the market against pre-existing competitors, giving input into arguments affecting what appear on labels or what makes it into payer formularies.

This does not come easy though. Each vendor is vying for position in this area and each year their data sources are being added to, linked together, tweaked and elaborated on so there is a constant need to adapt our approaches and uses of the data appropriately. Add to this the natural growth that is expected of the patients' data, then every little way that we interact with the data becomes significant and the scope of impact that we can have increases (we are almost seeing an exponential growth in some data sources). This mass of data that needs to be 'trawled' through in order to get answers can be daunting and generally only a well-planned and thought out way of dealing with it can yield accurate and timely results.

However, some people are going even further. There are ongoing efforts to build platforms that will allow patient identification, recruitment and clinical trial execution directly from hospitals' electronic medical records, all without having to go through CRFs i.e. mapped straight to SDTM format (Innovative Medicines Initiative¹, Electronic Health Records for Clinical Research²). Comparator treatment arms are being built using historical clinical trials data in order to save recruitment costs. If this is the case then why can't this be done using RWD? At the moment it is limited by the data types available, but every year the vendors are adding more robust and useful data and linkages, so it's only a matter of time before this becomes a reality.

This area of analytics is not new, but it is only now starting to come into its own. It is no longer enough for pharmaceutical companies to prove safety and efficacy of their drugs, but now they are expected to contextualize the use of their products in order to justify their place on the market as well as the benefit/risk portfolios as an ongoing regular activity.

This paper is not intended to be an introduction or training in RWD. Instead I hope that my personal perspective into the science and programming of RWD in the pharmaceutical industry has helped raise interest in this fascinating area. RWD is continually expanding and is even now changing the shape of how clinical trials are conducted and drugs are developed.

There is huge potential in RWD and its uses are almost limitless, so we need to grow with this and embrace it in order to maintain a competitive edge both for our companies and for ourselves as data scientists. RWD does not have all the answers however there is still a lot of power in it.

PhUSE 2013

REFERENCES (HEADER 1)

¹ <https://www.imi.europa.eu/web/portal/home>

² <http://www.ehr4cr.eu/>

RECOMMENDED READING (HEADER 1)

Health Care Data and SAS – M.Scerbo, C. Dickstein and A. Wilson, SAS Publishing

The Complete Guide to SAS Indexes – Michael A. Raithel, SAS Publishing

CONTACT INFORMATION

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

Rob Walls
F. Hoffmann-La Roche
Malzgasse 30
Basel 4070
Switzerland
Work Phone: (+41) 0616883036
Email: robert.walls@roche.com