The Medical Research Network

White Paper – By Dr Graham Wylie

September 2013

Utilising Big Data to Produce Faster and Cheaper Clinical Research

Recruitment and retention of patients remains the major challenge to ensuring trials run as fast as possible and give them the statistical power they are designed to produce. This is not just about getting better any more - but about stopping the deterioration of these metrics in modern clinical trials.

Focusing on recruitment, innovative companies have come up with several new techniques over the years which can be characterised as:

- Advertising - which increases the patient pool with access to the trial;
- Site support - finances etc - which reduces site workload in the site making them more able to recruit all the patients available in their pool;
- Home health care in trials – the most recent development, aimed at reducing the impact of the trial on the patient's life, which increases the rate of consent of patients in the pool.

Recruitment using internet based techniques represent the next frontier. So far the approach has had very mixed success. This appears to be because the creation of communities which offers the potential to interact with individual patients on a regular basis are based on trust and unbiased information which makes pharmaceutically sponsored sites poorly trusted sources, but also due to a lack of interest from Pharma companies who see that sort of interaction as breaching the firewall between them and patients, making the issues of confidentiality and coercion to participate in trials too close for comfort.

Alternatively, advertising on internet sites has had some success - presumably because it is similar to traditional media but has a much broader reach. There is no direct contact with patients allowing the threats of lack of confidentiality and coercion to recede. It is also a well understood process as we

have a long history of ethical committee approved advertising in traditional media to draw upon which is in fact very similar.

With the growth of Big Data, forms of data basing have also begun to proliferate. These seem to take three basic forms:

- Internet scraping - looking for data on published pages to create databases of possible patients and sites;
- Use of consumer data - from pharmacy services or similar data sets collected for other reasons;
- Brokerage - where a website serves to bring together and match sponsors of trials with patients interested in participation.

To this last set we can now add a new approach - that of *online* analysis. MRN is actively exploring this area in collaboration with PA Consulting Group and has created a new service called MRN-RAPID, based on "the social patient" concept from PA. This is a significant advance and starts to tap into the huge potential of the internet data that people publish about themselves, whilst still preserving confidentiality and avoiding any risk of coercion.

What is online analysis?

Online analysis is based on people sharing and seeking medical information to understand their condition, to access opinion on possible therapies and to access specialist information on research participation and results. All data analysed must be actively made public by the person publishing it. People put this data in their various status updates and on their profiles, discuss it and raise issues in their tweets and express it in their internet search patterns.

It is the scale of the data that makes it so useful. These social media systems reach billions of people worldwide, so even if a small number share this sort of information, that still amounts to hundreds of millions of people, making many statements each, in various media, in pretty much every country in the world, with considerable amounts of relevant data.

At this scale the data is indicative of the whole population, not just those on social media. As you expand the detail you require, the numbers will of course drop, but the quality of the pool of people you are defining as potential screening candidates rises until you have a very highly refined set of

possible patients for your trials. The data is also extremely current - not based on historical data beyond being a few weeks or at most months old (and you can define this criteria to exactly what you want). This gives you actual people you now want to target in a way that maintains confidentiality.

We are using the MRN-RAPID service as our example because it is the only service using this approach at the moment. So in our hands, the data generated after various degrees of filtering are analysed through a suite of 15 to 20 different software technologies. These can provide a huge amount of useful metrics about the population you are targeting as well as the actual patients in the most refined groups. The software has natural language capability, meaning it can provide insight into the attitude of the people publishing the data - for example if they are generally positive to the concept of taking part in a trial. It also allows removal automatically of publications which are perhaps jokes, selling products for the condition or not human (created automatically). Such volumes of data needs to be analysed automatically as much as possible, but the search patterns and analysis is also best guided by analysts with expertise in both utilising intelligence generated this way and those with trial knowledge.

The reports we need detail the physical and digital landscape for the people in the highly refined group. It is not possible to reach out to these people directly, and neither is this desirable as it stretches confidentiality to the maximum. However, we can learn a lot about them, and that allows us to put the option of participating in the trial in their likely field of activity on social media; essentially we want to put the information about our trial programmes in a place where they are most likely to find them.

To do this we analyse who influences them - what sites they look at for data, whose Twitter feeds they follow, where they publish their own data and how they interact with each other. By drawing up networks of connections we can identify those sites and Twitter feeds that will reach the highest number of these high quality screening candidates in the shortest time. We also want to know how this group search the internet generally and how to characterise them to target advertising based on their use of search engines. This, then, means we have real time metrics as to how many of the patients we are interested in are going to various sites where we can place adverts or information - all of which will of course be approved by ethical committees as required.

The next stage of the analysis is based on combining this data and other data sources to provide a geographical landscape for the potential patients. Since most people put the city they live in on their profiles, we know at that sort of level where these people live. At the least filtered level this gives a snap shot of the true demographics and distribution of the disease. In our Hepatitis C Virus (HCV) example this identified that in Italy there is a much higher incidence of the condition than expected from publications and text books. We do not know the reason for this, but we postulate that since Italy receives a very high number of immigrants from the northern areas of Africa where the disease is much more common this has increased the incidence in Italy. Drilling down to the highly filtered groups, we can see exactly what cities have the highest populations of socially active patients.

Using other data published about trials - covering the USA, Europe and Japan - we can then identify which cities have the highest incidence of competitive trials. Taking this info along with population numbers we can show a density of competitive trials per hundred thousand of the population and the social media activity in those cities, matching all three up to find the best place to put sites with the greatest numbers of patients and the lowest amount of competitive research.

The intent has to be to create a data set of high quality screening candidates, to know where to go to have the best chance of reaching out to them in the digital landscape and in the geographical landscape. This leads to the creation of marketing campaigns using Google ad words, Twitter hashtags that the potential patients follow, paid Twitter banners based on keywords used and banner advertising on influencing sites.

Advertising of this type is very cost effective. Google ad word campaigns are based on a very detailed set of search patterns and limited number of cities, which is therefore much cheaper than the scattergun approach otherwise favoured. In our example we were looking at reaching thousands of our target populations for around $15,000 per month.

Twitter is even cheaper. Having identified the top eight hashtags used by the highly refined target group, we can, for a few hundred dollars, create a set of tweets that can be reused throughout the recruitment period pointing patients to a screening website which will reach most of their data. We can also identify which of the influencer sites will take banner ads that cost just a few hundred dollars to create and a few thousand to post per month, yet will hit huge numbers of high quality, highly refined potential patients.

The cost efficiency introduced is very high. As for site placement - the aim here is to have fewer sites by eliminating those who are not likely to recruit patients because they have too much trial competition and too few patients. You can further reduce site numbers by recruiting through the internet and referring patients to the most well resourced and effective sites.

Why does it represent an advance in this field?

Advantages of this approach are:

- SCALE - we can look at billions of patients and their data points at the start of our analysis - even the largest big data databases in this field cover no more than a couple of hundred million patients.

- CURRENCY - the data is not historical beyond a few weeks or months. This means there will be very limited 'false positives' due to patients being reported who might have fit criteria in the past but no longer do so, or sites that might have had the interest, expertise and capacity to do trials in the past but no longer do so.

- GEOGRAPHICAL REACH - the internet and social media is ubiquitous in most countries in the world, whereas other systems often focus on just a few or indeed just on the USA. This data is even available from government controlled systems such as those in China.

The bottom line is:

- Fewer sites should fail to recruit;
- Advertising should be much better targeted on actual metrics and be much cheaper;
- High quality potential patients can be reached and informed of the trial in a short period of time;
- All of which leads to faster recruitment and reduced costs.

How are we making it work?

We have run a case history on HCV. We created a set of criteria against which to analyse the data in social media - looking back over six months. This generated a data set of 500,000 conversations.

We then refined that data set down to several thousand using the criteria of age, sex, duration of disease, genotype of HCV, previous therapies (successes and failures), disease severity - all of which are often published by people online.

We were able to identify 18 key influencers - 10 on the internet and eight in Twitter. Of these about half would take banner advertising.

Site placement identified several cities in the US and other countries where the incidence of social media activity was high, the population was high, but the competitive trial presence was low. We also identified the opposite - cities that should not be used.

From this we were able to identify where to advertise to reach this high quality screening pool - what sites they go to, whether they take advertising, how to micro focus your Facebook and Google ad words advertising, and how to publish via Twitter using hashtags and keywords to get maximum target audience exposure . This is not based on theory, but on actual real time metrics.

This case history has not progressed to the advertising campaigns, so the final outcomes are still to be determined.

What is the future potential?

This system is only just coming alive. As yet we are just asking simple questions which, despite being powerful in terms of enhancing recruitment, are just scratching the surface of what can be done. There remains a lot of development to undertake yet to realise the full potential. We *could* have asked:-

More about sites:
What sites are there in the target cities and are they interested in trials?
Do the sites have a good access to patients?
Are they 'influencing' the local population? Do they communicate with patients well?
Do they have an opinion leader in them, do they have trial history, etc?

More about patients:
We could ask what patients think of elements of trial design; what attracts them and what puts them off?

More about what trial design:

We could track how people's posts change related to specific topics raised on Twitter or through influencers, or we could literally ask market research type questions about willingness to participate in clinical trials. We could also find out how satisfied or otherwise patients are with trials they are already in or have been in the past to make sure new designs are as attractive as possible.

We could effectively poll people as to their concerns about protocol design by asking influencers to discuss various design elements such as comparative agent, duration of therapy, willingness to travel to sites and any other relevant topic. We can then watch the published data and analyse what they are saying on the topics.

More about screening tools:

Where to put a screening website for the best response?

Where might we get the best click through due to trust and security?

Conclusion

The internet age is well and truly with us. Patients are already publishing huge amounts of information about themselves, including their illnesses, their medications, their trial experiences etc. The data is out there, and the industry must not ignore it, or we will fail to adapt to patients being better informed, more inquisitive, more networked and more in control. Using it has to be done carefully and respectfully, in ways people expect their data to be used in the new digital age. Our traditional methods - crumbling already - will start to fail more comprehensively, and the prize (faster and cheaper research) will go to the company that acts first to reverse the trends in its R&D programmes. Social media and internet based analysis of big data using these online techniques are making big databases redundant as they look more and more out of date, unwieldy and lacking in scale and reach. Using these tools whilst still respecting confidentiality and avoiding taints of coercion is available here and now and will make an excellent solution to our long term problems of recruitment.

How to find out more:

If you want to know more about the subject drop us a line or look us up on the web at www.themrn.co.uk and LinkedIn at http://www.linkedin.com/company/medical-research-network.

The Medical Research Network

UK Address:

Medical Research Network Limited
Talon House
Presley Way
Milton Keynes
Buckinghamshire
MK8 0ES

Tel: +44 (0) 1908 261 153
Email: Enquiries@themrn.co.uk

USA Address:

Medical Research Network Inc.
4030 Wake Forest Road – Suite 300
Raleigh, NC 27609
USA

Tel: +1 (919) 719 7222
Email: Enquiries@themrn.co.uk