ELSEVIER



# Developing timely insights into comparative effectiveness research with a text-mining pipeline

Q1 Meiping Chang[1], Man Chang[2], Jane Z. Reed[3], David Milward[3], Jinghai James Xu[4] and Wendy D. Cornell[5]

[1] Informatics IT, Merck & Co., Inc., Boston, MA, USA
[2] Linguamatics, Westborough, MA, USA
[3] Linguamatics, Cambridge, UK
[4] Corporate Development, Merck & Co., Inc., Kenilworth, NJ, USA
[5] Informatics IT, Merck & Co., Inc., Rahway, NJ, USA

**Comparative effectiveness research (CER) provides evidence for the relative effectiveness and risks of different treatment options and informs decisions made by healthcare providers, payers, and pharmaceutical companies. CER data come from retrospective analyses as well as prospective clinical trials. Here, we describe the development of a text-mining pipeline based on natural language processing (NLP) that extracts key information from three different trial data sources: NIH ClinicalTrials.gov, WHO International Clinical Trials Registry Platform (ICTRP), and Citeline Trialtrove. The pipeline leverages tailored terminologies to produce an integrated and structured output, capturing any trials in which pharmaceutical products of interest are compared with another therapy. The timely information alerts generated by this system provide the earliest and most complete picture of emerging clinical research.**

## Introduction

Q3 Clinical trials form the cornerstone of evidence-based medicine and are essential to establishing the safety and efficacy of new drugs. The relative merits of different therapies for the same disease are also of interest and, with the American Recovery and Reinvestment Act of 2009, the US Government is investing over US$1 billion in the support of CER [1–3]. According to the US Department of Health & Human Services (HHS) Agency for Healthcare Research and Quality, CER is 'designed to inform health-care decisions by providing evidence on the effectiveness, benefits, and harms of different treatment options' (http://effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/). Widespread adoption of CER has the potential to drastically change healthcare by increasing quality while reducing otherwise rising costs, guiding payers in formulary design and physicians and patients in personalized treatment plans optimized for specific disease conditions and patient populations. Pharmaceutical companies use CER to identify and address market opportunities and risks related to existing products and those in development [4,5].

There are several methodologies available to address CER questions, including (i) interventional clinical trials (usually randomized controlled trials or RCTs) [6]; (ii) prospective observational real-world evidence studies (routine clinical setting) [7]; (iii) retrospective observational real-world evidence studies (based on electronic medical records, claims information, patient registries, and patient surveys) [8,9]; and (v) systematic review and meta-analysis of data from multiple clinical trials or studies [10,11].

The first option of interventional RCTs is considered to be the gold standard for evidence-based research, because the statistically rigorous design leads to internal validity [12]. Recommendations have been made to develop new paradigms for RCTs, for example, by including more arms [13] or incorporating the external validity [14] of the pragmatic observational approaches with their fewer exclusion criteria and real-world accommodations [15–17].

Modern clinical research is conducted under principles established in 1964 by the Declaration of Helsinki from the World Medical Association (WMA) that acknowledged the need to carry

*Corresponding author:* Chang, M. (meiping.chang@merck.com)


http://dx.doi.org/10.1016/j.drudis.2016.01.012


Please cite this article in press as: Chang, M. et al. Developing timely insights into comparative effectiveness research with a text-mining pipeline, Drug Discov Today (2016), http://dx.doi.org/10.1016/j.drudis.2016.01.012

out medical research on humans and established ethical obligations, such as informed consent and the publication and dissemination of results [18]. 'Every research study involving human subjects must be registered in a publicly accessible database before recruitment of the first subject' and 'researchers, authors, sponsors, editors and publishers all have ethical obligations with regard to the publication and dissemination of the results of research' [18]. The US Food and Drug Administration Modernization Act (FDAMA) of 1997 [19] mandated the establishment of a clinical trial registry to capture trials conducted under investigational new drug (IND) applications, resulting in 2000 in ClinicalTrials.gov, from the National Library of Medicine (NLM) of the National Institute of Health (NIH) [20]. The FDA Amendments Act of 2007 (FDAAA) expanded the scope of trials requiring registration and the required information and also led to the establishment of a trial results database in 2008 [21]. The requirement for reporting of adverse events was added in 2009.

In 2004, the International Committee of Medical Journal Editors (ICMJE) set as a condition for publication of Phase III trials their registration on or before the date of first subject enrollment in an electronically searchable public trials registry managed by a not-for-profit organization, free of charge, and open to all prospective registrants [22,23]. The requirement was amended to include completion of 20 specified data fields defined by the World Health Organization (WHO) and adopted by ICMJE. They noted, 'A complete registry of trials would be a fitting way to thank the thousands of participants who have placed themselves at risk by volunteering for clinical trials. They deserve to know that the information that accrues from their altruism is part of the public record, where it is available to guide decisions about patient care, and deserve to know that decisions about their care rest on all of the evidence, not just the trials that authors decided to report and that journal editors decided to publish' [22,23].

At the time the ICMJE policy was announced, ClinicalTrials.gov was the only registry that met the set requirements, but soon after, in 2006, the WHO International Clinical Trials Registry Platform (ICTRP) [24] was established and most of the primary national registries to which it offers access also meet the ICMJE criteria [23]. As a result, a considerable amount of information is available in both ClinicalTrials.gov and ICTRP [25]; however, compliance has been an issue and has been the subject of multiple recent publications with respect to both trial registration [26,27] and posting of results [28]. Other studies have taken issue with the slow time to publication [29] or the lack of publication [30] in the scientific literature. Finally, other studies have identified discrepancies between the registered and published forms of particular trials, for example with respect to stated primary outcomes [31,32]. These findings have led to recent calls from both the WHO and NIH to improve compliance and data sharing [33–35]. At the time of writing this article, two policy changes were under review by the US Department of HHS; the first proposed expanding the scope of applicable clinical trials to include those of unapproved drugs, biologics, or devices, and the second proposed enforcements for compliance with registration and reporting for NIH-funded trials (https://www.federalregister.gov/articles/2014/11/21/2014-26197/clinical-trials-registration-and-results-submission; https://www.federalregister.gov/articles/2015/02/13/2015-02994/announcement-of-a-draft-nih-policy-on-dissemination-of-nih-funded-clinical-trial-information). As of 2014, results reporting is now also required for the European Clinical Trials Register (EU-CT), one of the registries accessible through ICTRP [36].

Even with these compliance issues, the clinical trial registries contain much valuable information that can be used in trial site selection, trial design [37,38], and assessment of patient inclusion and exclusion criteria for particular therapeutic areas [39,40]. Other applications include the identification of competitive intelligence or information on licensing opportunities [41] and the assessment of trial characteristics for different disease areas or patient or geographic populations [42–47]. However, there are issues regarding effective search; for example, search vocabularies are incomplete and much of the information within records is unstructured and not easily extracted using the registry-supplied search interfaces. Here, we review the search provision for ClinicalTrials.gov and WHO ICTRP and describe a workflow that uses a text analytics approach to search and synthesize comparative effectiveness data from three different sources in a more effective and efficient process.

## Native Clinical Trial Registry search interfaces

ClinicalTrials.gov (https://www.clinicaltrials.gov/) and the ICTRP portal (http://apps.who.int/trialsearch/Default.aspx) each offer interfaces supporting both simple 'Google' (https://www.google.com)-type search where keywords related to indication and location are entered into a single box as well as more advanced search capabilities targeting specific fields with drop-down or free text, such as recruitment, study type, study results, conditions, interventions, titles, and so on. Both sites support Boolean operators and recognize at least some synonyms, with the ClinicalTrials.gov site appearing to have a more reliable and comprehensive list. For example, a search on ClinicalTrials.gov for Merck's diabetes blockbuster drug yielded the same number of hits regardless of which of five synonyms was used: 'Januvia'®, 'Sitagliptin', 'MK-0431', 'MK0431', or 'MK 0431'. By contrast, the same searches on the ICTRP portal led to three different results sets depending on whether the brand name, generic name, or an orthographic variant of the company code was used. In other words, ClinicalTrials.gov recognized Januvia®, Sitagliptin, and MK0431 to be synonyms, but ICTRP did not. Both sites listed the synonyms used in the search result, revealing in this case that 'Xelevia'®, a brand name for this drug used in Europe, is yet another synonym included in the ClinicalTrials.gov search (accessed 6 August 2015). By contrast, both ClinicalTrials.gov and ICTRP recognized two synonyms for another blockbuster: 'Lipitor'® (aka 'atorvastatin'). These differences are important because the performance of any NLP approach depends heavily on the terminologies used. If relevant synonyms are not built into the workflow, they must be input manually by the user.

ClinicalTrials.gov offers a third 'expert' search option (available from the search modification page) where search terms for all fields appear in a large single box accommodating easy entry and editing of multiple 'cross-terms' and subqueries in a single search. A specialized preformulated 'multi-query' for CER is available (https://www.nlm.nih.gov/nichsr/cer/cerqueries.html) and can be modified as desired, for example, to add a drug name, eliminate certain study or intervention types, expand the phases included, or add dates or sponsors.





Please cite this article in press as: Chang, M. et al. Developing timely insights into comparative effectiveness research with a text-mining pipeline, Drug Discov Today (2016), http://dx.doi.org/10.1016/j.drudis.2016.01.012