Published in final edited form as:

Trends Pharmacol Sci. 2015 November ; 36(11): 706–709. doi:10.1016/j.tips.2015.08.007.

Optimizing Clinical Research Participant Selection with Informatics

Chunhua Weng, PhD

Department of Biomedical Informatics, Columbia University, 622 W 168 Street, PH-20, Room 407, New York City, NY 10032

Chunhua Weng: cw2384@columbia.edu

Abstract

Clinical research participants are often not reflective of the real-world patients due to overly restrictive eligibility criteria. Meanwhile, unselected participants introduce confounding factors and reduce research efficiency. Biomedical Informatics, especially Big Data increasingly made available from electronic health records, offers promising aids to optimize research participant selection through data-driven transparency.

Keywords

Selection bias; Clinical trial; Informatics; Patient-Centered Care

The trade-off between generalizability and internal validity for clinical studies

Clinical studies are fundamental for translating breakthroughs in basic biomedical sciences into knowledge that can benefit clinical practice and ultimately human health. When selecting study participants, researchers must consider scientific, ethical, regulatory, and safety requirements and translate these into unambiguous eligibility criteria [1]. However, overly restrictive participant selection has compromised study generalizability, severely impaired the cost-benefit ratio of clinical studies, and contributed to the difficulty in implementing and disseminating study results to real-world patients across many disease domains [2–4]. Moreover, the lack of study generalizability largely remains undiscovered until after study publication or during systematic reviews. There is no scalable and proactive approach to predicting a priori generalizability during clinical study designs.

Clinical research eligibility criteria: the hidden key to balancing the trade-off

Clinical research eligibility criteria play an essential role in clinical and translational research. They influence study generalizability by defining the characteristics of the target populations of clinical studies, which are interpreted, implemented, and adapted by different stakeholders at various phases in the clinical research life cycle. After being defined by investigators, criteria are used and interpreted by research coordinators for screening and recruitment. Query analysts, and even research volunteers themselves, each possess different

decision support needs for using the eligibility criteria for screening. Later, the criteria are summarized in meta-analyses for developing clinical practice guidelines and, eventually, interpreted by physicians to screen patients for evidence-based care. The quality of eligibility criteria directly affects recruitment, results dissemination, and evidence synthesis. Poorly designed eligibility criteria have been reported to slow down recruitment; cause early study termination; increase costs; impair study generalizability; and either exclude patients who might benefit from experimental therapies or, conversely, threaten patient safety by leading to post-marketing adverse drug effects.

What is wrong with the current practice for eligibility criteria design

Clinical research eligibility criteria have been criticized for their vagueness, ambiguity [5], complexity [6], overly restrictive nature, lack of patient-centeredness [7], and lack of computational capability and interoperability across studies or with other data sources [8]. Unfortunately, few resources exist for helping clinical investigators discover potential patient selection problems and make better eligibility criteria choices. Rarely is the rationale behind eligibility criteria choices provided explicitly, adding to the difficulty in problem detection and correction. These problems stem from clinical researchers' limited knowledge of the etiology or comorbidities of many diseases and researchers' lack of precise understanding and characterization of the real-world patients. Consequently the existing popular practice for eligibility criteria definition is suboptimal. Many clinical researchers rely upon past experience and knowledge to conceptualize and select population subgroups into clinical studies. However, this participant selection process is at best subjective and unsystematic [9]. Study designers often copy and paste eligibility criteria from related clinical research protocols with only slight adaptations [10], reinforcing converging selection of certain population subgroups and collectively exacerbating health disparities among population subgroups, which are studied either rarely or excessively. Moreover, eligibility criteria are often defined through trial and error, a process that requires many protocol amendments. There is an unmet need for early protocol feasibility assessment and cost-effectiveness analysis of individual eligibility criteria [7].

A new opportunity for transforming participant selection using Big Data

The imperative need to strike the balance between generalizability and internal validity of a clinical study is essentially an optimization problem that can benefit from data-driven transparency. The recent and continual burgeoning of electronic health records, clinical data warehouses, and clinical data networks have made available enormous amount of electronic patient data. Notable examples include the national multi-site PCORnet Clinical Data Research Networks (CDRNs) (<http://www.pcornet.org/clinical-data-research-networks/>) and the international collaborative network for Observational Health Data Sciences and Informatics (OHDSI) (<http://www.ohdsi.org>). Such data infrastructures enable scaled analytics for patient modeling and population profiling. In addition, clinical study design information is increasingly available publicly, especially through the public ClinicalTrials.gov, to improve both the transparency of clinical research and the public's trust in such research. These data resources offer an unprecedented opportunity to explore patient-centered, knowledge-based, and data-driven eligibility criteria design through

computational modeling of population subgroups and participant selection for clinical studies. Figure 1 illustrates our vision towards this goal.

In this vision, existing electronic clinical, environmental, or genetic patient data are integrated or federated and then summarized to develop a digital representation of the real-world population, which is then used for deep phenotyping and subgroup modeling. Common eligibility variables and their usage trends in clinical studies are automatically mined from public clinical research summaries, such as those in ClinicalTrials.gov, to inform clinical research knowledge reuse and parameterization for subgroup modeling. Outcome data for eligibility criteria, as measured by the publication records of each clinical study, can be extracted from PubMed. The subgroup characteristics are then automatically summarized to generate human-understandable and machine-computable eligibility criteria text and presented to key stakeholders of clinical research studies, including sponsors, patients, clinicians, and the researchers themselves, who can consider “whom are frequently excluded in existing studies” and “whom should be studied” when they attempt to balance external validity – i.e., generalizability – and internal validity based on group consensus.

This model for participant selection is expected to (1) enable *in silico* assessment of early feasibility and a priori generalizability, and optimization of eligibility criteria at the study level; (2) increase transparency for clinical research participant selection and detect and bridge evidence gaps at the systems level; (3) facilitate shared decision-making for participant selection among key clinical research stakeholders; (4) enable flexible and continuous modification of eligibility criteria based on real-time data-driven feedback; and (5) ultimately improve patient-centeredness of clinical studies and thus reduce health disparities.

Informatics as Enabler

Informatics is essential to achieve this vision. The science of informatics drives innovation that defines future approaches to information and knowledge management in biomedical research, clinical care, and public health (www.amia.org). Advances in biomedical informatics, especially in natural language processing, electronic health records-based data reuse, and visual analytics, have enabled the development efforts necessary to achieve this vision.

Advanced natural language processing systems can transform large amounts of text from PubMed or ClinicalTrials.gov into discrete and computable information for aggregate analysis of clinical research design patterns for participant selection. For example, these systems can be used to mine all cancer studies to identify the most frequently used eligibility criteria for clinical studies on cancer patients. The visual aggregate analysis system VITTA allows users to interrogate ClinicalTrials.gov for frequently used medical concepts in eligibility criteria in any disease domain and their common value ranges [11].

Research on electronic health records has increased our understanding of their value as well as their limitations and has made available scalable approaches to modeling patients, clinical phenotypes, health outcomes, and population characterization. Linking public clinical trial knowledge and electronic patient data, we recently compared the value distributions for age

and A1c for about 20,000 Type 2 diabetes patients to the value distributions of the age and A1c eligibility criteria in 1,761 Type 2 diabetes trials and confirmed the known fact that the target populations in diabetes trials tend to be younger and sicker than real-world diabetes patients [12]. These results were replicated using a national survey of population health database, NHANES, to avoid potential bias in an individual institution's clinical data [13]. These studies proved the feasibility of data-driven *priori* generalizability assessment so that in the future such assessments do not have to wait until the completion and publication of clinical studies. The data-driven methods are also more scalable and cost-effective than existing manual methods.

Challenges and Recommendations

Several research challenges must be overcome in order to achieve the vision of data-driven participant selection. In order to support data-driven generalizability assessment for a clinical study, it is necessary to model all possible eligibility criteria variables and all possible values, especially for every numerical eligibility variable. Therefore, the extremely high dimensionality involved in population subgroup modeling requires more sophisticated models than are currently available. This also necessitates interdisciplinary collaboration between informatics and statistics. In addition, sampling bias and data incompleteness are two major barriers to reusing existing electronic patient data to understand the real-world patients [14, 15]. These electronic data need to be supplemented with patient self-reported outcomes, genetic or environmental data, public records of clinical study outcomes, and other electronic data that can be semantically linked to profile the clinical research design patterns and outcomes. Achieving the semantic interoperability of isolated data sources is another important yet difficult task. Finally, optimization simulation experiments are still rare for eligibility criteria design and need substantial development. A socio-technical approach is necessary to capture the preferences of clinical research stakeholders and then apply an optimization model to these preferences. Policies will be needed to promote a new culture for data-driven eligibility criteria optimization. Health literacy barriers from research volunteers can prevent research volunteers from effective participation in the shared decision-making for eligibility criteria definition. The latest advances in the field of natural language generation should be leveraged to automatically translate population subgroup characteristics into comprehensible while computable eligibility criteria, ideally based on well-adopted standards for clinical research common data elements.

Conclusions

The increasing availability of big clinical data offers new and promising opportunities to supplement domain expertise with data-driven optimization of clinical research eligibility criteria through iterative *in silico* *a priori* generalizability assessment.

Acknowledgments

This research was funded by National Library of Medicine grant R01 LM009886 (PI: Weng).

References

1. Kim ES, Bernstein D, Hilsenbeck SG, Chung CH, Dicker AP, Ersek JL, et al. Modernizing Eligibility Criteria for Molecularly Driven Trials. *J Clin Oncol*. 2015;10.1200/JCO.2015.62.1854
2. Masoudi FA, Havranek EP, Wolfe P, Gross CP, Rathore SS, Steiner JF, et al. Most hospitalized older persons do not meet the enrollment criteria for clinical trials in heart failure. *American heart journal*. 2003; 146(2):250–7. Epub 2003/08/02. 10.1016/s0002-8703(03)00189-3 [PubMed: 12891192]
3. Schoenmaker N, Van Gool WA. The age gap between patients in clinical studies and in the general population: a pitfall for dementia research. *The Lancet Neurology*. 2004; 3(10):627–30. Epub 2004/09/24. 10.1016/s1474-4422(04)00884-1 [PubMed: 15380160]
4. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA*. 2007; 297(11):1233–40.10.1001/jama.297.11.1233 [PubMed: 17374817]
5. Musen MA, Rohn JA, Fagan LM, Shortliffe EH. Knowledge engineering for a clinical trial advice system: uncovering errors in protocol specification. *Bull Cancer*. 1987; 74(3):291–6. [PubMed: 3620734]
6. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*. 2010; 2010:46–50. Epub 2011/02/25. [PubMed: 21347148]
7. Sharma NS. Patient centric approach for clinical trials: Current trend and new opportunities. *Perspect Clin Res*. 2015; 6(3):134–8.10.4103/2229-3485.159936 [PubMed: 26229748]
8. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. *Journal of biomedical informatics*. 2010; 43(3):451–67.10.1016/j.jbi.2009.12.004 [PubMed: 20034594]
9. Rubin DL, Gennari J, Musen MA. Knowledge representation and tool support for critiquing clinical trial protocols. *Proceedings of the AMIA Symposium*. 2000:724–8. Epub 2000/11/18.
10. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. *Journal of biomedical informatics*. 2014; 52:112–20. Epub 2014/02/06. 10.1016/j.jbi.2014.01.009 [PubMed: 24496068]
11. He Z, Carini S, Sim I, Weng C. Visual aggregate analysis of eligibility features of clinical trials. *Journal of biomedical informatics*. 2015; 54:241–55.10.1016/j.jbi.2015.01.005 [PubMed: 25615940]
12. Weng C, Li Y, Ryan P, Zhang Y, Liu F, Gao J, et al. A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Applied clinical informatics*. 2014; 5(2):463–79. Epub 2014/07/16. 10.4338/aci-2013-12-ra-0105 [PubMed: 25024761]
13. He, Z.; Wang, S.; Borhanian, E.; Weng, C., editors. Assessing the Collective Population Representativeness of Related Type 2 Diabetes Trials by Combining Multiple Public Data Resources; *Proc of MedInfo'2015*; Sao Paulo, Brazil. 19–23 August; 2015. accepted
14. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association : JAMIA*. 2013; 20(1):144–51.10.1136/amiajnl-2011-000681 [PubMed: 22733976]
15. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annu Symp Proc*. 2013; 2013:1472–7. [PubMed: 24551421]

Figure 1.

A vision for data-driven, knowledge-based, and patient-centered participant selection for clinical and translational research. In this vision, participant selection starts with population characterization and subgroup modeling, which is followed by evidence gap analysis. Key stakeholders of clinical research will be informed with real-world patient needs and reusable eligibility criteria knowledge to decide “whom should be studied.”