

# A Methodology for Quantitative Measurement of Quality and Comprehensiveness of a Research Data Repository

Vojtech Huser, MD PhD

Marshfield Clinic, Biomedical Informatics Research Center, Marshfield, WI, USA

## Background

With the existence of research over federated repositories, it is desirable to utilize high quality integrated data repositories (IDRs). IDR can be defined as a data warehouse optimized for research purposes rather than clinical care, which contains clinical, administrative, trial, and -omics data [1]. In this work, we focus on the quality of clinical and administrative components of IDRs. There is no standardized methodology which could quantitatively evaluate the quality of an IDR (e.g., “Does a given IDR have at least 2000 adult diabetic patients (type 1) with complete pediatric history?”). With the increased interest in research of existing data (such as comparative effectiveness research) and increasing number of institutions with an comprehensive IDRs, it is important to have a mechanism for quickly selecting quality IDRs.

## Methods

Our poster will present a set of IDR quality measures which can compare IDRs in size and completeness. We considered the following criteria for a good measure: the measure is intuitive to interpreted (such as count of patients); facilitates monitoring improvement; and does not place any arbitrary value on individual measure components (e.g., value of 10

years of medication history vs. 10 years of weight/height history).

Our methodology proposes a hierarchy of definitions of minimum EHR elements (see Table 1 for examples) and uses a simple count of each level to quantitatively evaluate an IDR.

## Results

We have applied our methodology to an IDR at Marshfield Clinic (see Table 1). To facilitate evaluations at other institutions, we have created an ANSI-SQL script which can compute all current measures in a single execution.

## Conclusion

Our evaluation methodology provides a quick way to compare IDRs at different institutions. It can be applied to institutions contributing to a virtual warehouse. Our goal was to arrive at a pragmatic set of measures operating on an easy-to-implement event schema. The limitations are: focus on general research idea and event types and criteria included in some level definitions. We plan to conduct a Delphi study involving informatics experts to arrive at an improved consensus set of measures.

Table 1: IDR SnapShot v1.0 measures and results (columns: Level, definition (counts of), result)

Level	Definition	Result
<b>General dimension</b>		
G1	all events	0.5 B
G2	unique patients	1.7 M
<b>Data Dimension</b>		
D1	count of patients with at least one diagnosis and one laboratory result	0.755 M
D2	same as D1 plus at least one clinical report	0.720 M
D3	same as D1 plus at least one prescription	0.429 M
D4	D2 and D3 combined	0.429 M
<b>Lifetime dimension</b>		
L1	count of patients with at least one pediatric event	0.261 M
L2	count of patients with at least one pediatric event and one adult event	0.134 M

## References:

[1] CTSA: Informatics Data Repositories Best Practices Symposium: Integrated Data Repository Survey (October, 2008) available at: [http://www.ctsaweb.org/uploadedfiles/CTSA\\_IDRSurveyResults20081017.ppt](http://www.ctsaweb.org/uploadedfiles/CTSA_IDRSurveyResults20081017.ppt) [accessed Oct 27th, 2009]