What Records Should VA Keep About Phenotypes?

Andrew Zimolzak, MD, MMSc

March 27, 2017

Like essentially all VA studies that rely heavily on the Corporate Data Warehouse, MVP must deal with the fact that data elements are incompletely harmonized across the 130+ VA Medical Centers or "stations,"[1] although messy data is certainly not unique to VA.[2] Even when stations are forced to use the same coding system nationwide (e.g. ICD diagnosis codes), we must perform further work to determine the reliability of any combination of data elements. *Phenotype* is the term we use for a set of harmonized data elements, or for a higher-order concept made of many such "building blocks." When a subject matter expert (SME), analyst, statistician, and/or data pull engineer create a phenotype, the result can be: SQL code that represents the SME's rules, a public database table that represents the enumerated list of included/excluded data items, or even a statistical model (comprising SAS or R or other code, plus coefficients) of the likelihood of a patient having the given phenotype.[3]

[1] "[D]ata aggregation across the VHA is highly problematic, and data validity is often impossible to verify." *N Engl J Med* 373:1693–1695

[2] Example hemoglobin A1c units: "%, %HB, % A1C, MG/DL, G/DL, NULL, Blank, U." *Pharmacoepidemiol Drug Saf.* 23(6):609–18.

[3] Example phenotypes, which happen to be requirements from *non*-MVP studies: Failed first-line platinum-containing chemotherapy. Urgent coronary revascularization (completed or attempted) because of unstable angina. Erectile dysfunction, defined as first prescription for PDE5 inhibitor or referral for ED.

What we can't do today

Ideally, a robust phenotype library should exist, serving several functions. I am deliberately steering clear of discussing *how* to serve each function and would rather leave that discussion for the group.

(1) The library should make the products of each SME/data review process *available for viewing* by study coordinators, either by searching, or by browsing an overview of all available phenotypes; and these products should be *available for immediate use* by study analysts writing further analytic code. (2) The *code* used in the process of deriving a phenotype should be preserved, in part because the scientific validity of a study's conclusions rests in part on the phenotype code. (3) Phenotype and code *metadata* (such as responsible SME and data extractors, and date of creation) should be clearly noted, to allow study staff to trust a phenotype, and to allow phenotyping core personnel to renew elements over time. (4) The library should *document* the rationale for decisions made by SME, analyst, etc., and the functioning of code, and it should allow collaborative editing of this documentation. Documentation will typically be in sentences and paragraphs.[4]

[4] E.g. "[S]hould the ETL process be interrupted by any kind of database, host machine or connection issue. . . , in order to protect data and avoid partial load the ETL will not re-execute itself and. . . ." Courtesy of Oleg Soloviev.

A central idea of functions (1)–(4) above is that the methods of science should be available and documented well enough to be reproducible—in other words, "show your work."[5] Often for our type of science, *code is method*.[6]

Wish list (a.k.a. thinking big)

Practical: For some adjudications that SMEs perform (e.g. laboratory), we could develop a system that infers/learns a SME's internal mental model,[7] from the SME answering "keep/discard" to database elements, with only slight further effort from the SME. For the very most popular phenotypes, we could advertise database counts to study coordinators, updated monthly perhaps. Phenotypes could be linked back to specific clauses in a given version of the research study protocol; any amendment to a critical part of the protocol would trigger a need for an amendment to the corresponding data process.

Scientific: I personally imagine keeping track of all analyses run on the data set, starting even at the early exploratory stage of phenotyping. I would imagine the same record-keeping attitude that one might have toward a laboratory notebook in a basic bioscience lab, or toward a patent book.[8] When study results are published, will it be possible to run the whole analysis with one click (at least within VA, where the data resides)? If not, *why not?*

VA Collaboration: VA would benefit greatly if the library were available across VA nationally: if it made phenotypes *available for reuse*, for the purposes of other projects. Phenotyping is performed very commonly not only for MVP studies, but for large VA Cooperative Studies Program (CSP) trials—coordinated in Boston and at the other centers. Furthermore, smaller, non-CSP research projects at single VA Medical Centers would benefit, as would non-research efforts (quality and business related).

From my personal perspective, MVP communicates well about phenotyping with certain CSP and other studies based in Boston. Those communications are relatively informal, however. We have made in my opinion somewhat slow progress synchronizing efforts with VA groups apart from the above. This may be because it is a time-consuming proposition to share a large amount of history and accumulated experience from one group to another and vice versa, to make processes transparent enough for others to trust, and then to decide on a common direction. Furthermore, after a "merger," both sides are likely to see that some of their work (sunk costs) does not get incorporated in the final product.

[5] "[O]ur policy now mandates that when code is central to reaching a paper's conclusions, we require a statement describing whether that code is available and setting out any restrictions on accessibility." *Nature* 514:536, 2014-10-30

[6] sciencecodemanifesto.org

[7] E.g. "serum sodium is where LabChemTestName matches these 3 strings, or where LOINC is one of these 5 codes, except in these 20 cases. . . ."

[8] "Just as with biological materials, such code should be made available alongside results obtained using it. . . . Because software support is far more burdensome than delivery of materials such as plasmids, this code may be accompanied by a disclaimer. . . ." *Nature Methods* 11 (3): 211. March, 2014.