CONFIDENTIALITY
COALITION

Principles on De-identification and Use/Disclosure of De-identified Data Sets

HIPAA covered entities (CEs) and business associates (BAs) and individuals and entities that use health data for research and healthcare operations purposes are confident that the standards for de-identification of protected health information (PHI) listed in 45 CFR Subtitle A § 164.514 are, and have been, effective in allowing the use and dissemination of health data for critically important scientific and patient care improvement purposes while effectively protecting the privacy of individuals. However, we recognize that techniques for re-identifying 'anonymized' data have advanced, that many people are concerned about the potential misuse of health information, and that medical identity theft is growing.

The Confidentiality Coalition strongly believes that the forthcoming Guidance on "how best to implement the requirements for" de-identification under § 164.514 should not suggest practices or contain recommendations that would, purposely or inadvertently, diminish the availability or utility of de-identified health data for research and healthcare operations purposes. Whether the objective is comparative effectiveness research (CER) or health care quality improvement, treatment benefit analysis, or the sharing of scientific knowledge via case reports, de-identified data is crucial.

As organizations committed to responsible use and disclosure of de-identified data, we urge HHS to consider recommending certain 'best practices,' including:

- In regard to de-identified data sets that will be disclosed outside of the entity that collected and aggregated the component data, or that will be used and/or may be re-disclosed by an individual or organization that received a de-identified data set, we recommend the use of contract language that prohibits the recipient of the de-identified data set from attempting to re-identify individuals in the data set;

- In regard to de-identified data, we recommend that reasonable and appropriate safeguards be in place to protect the data.

Recognizing that HHS can only provide Guidance within the scope of its regulatory authority, the Confidentiality Coalition is concerned about the potential misuse involving publicly available "reference" information databases, such as hospital discharge databases, or motor vehicle databases, administered by the states or other entities. While such databases may provide societal value in regard to the transparency and accountability of 'public' or 'government' functions, they also can be used by adversaries to match key attributes of disparate data sets and thus potentially access identifiable health information. As law professor Paul Ohm notes in *Broken Promises of Privacy: Responding to the*

Surprising Failure of Anonymization, “It is no coincidence that every case study presented involved the public release [emphasis added] of anonymized data.”[1]

We believe the Department could provide a genuine public service by articulating that any data set containing de-identified health data that is publicly displayed, such as on the Internet, without a data use or other contractual agreement in place, should always be “scrubbed” of data elements that might reasonably be foreseen as facilitating re-identification of any individual. This could serve to extend the standards of § 164.514 to all large data sets that contain health data.

An alternative, or perhaps complementary, recommendation would suggest that any de-identified health information data set that is to be made publicly available (whether by a State or the CDC, for example) require a recipient seeking access to the data to agree to a contractual obligation (via a click-through or other online method, for instance) requiring that reasonable safeguards be in place and explicitly prohibiting re-identification of or contact of individuals whose information is contained in the data set.

In summary, the Confidentiality Coalition urges the Department to encourage the use of the best practices noted above as the basis for enabling the flow of de-identified data to advance scientific breakthroughs and improve healthcare quality. Important research and policy efforts aimed at improving patient safety, quality and outcomes could be at risk if the very information needed to implement these efforts is unduly restricted as a result of the forthcoming Guidance on the HIPAA Privacy Rule’s De-identification Standard.

[1] Ohm’s paper describes the ‘easy re-identification’ of three anonymized databases: AOL search queries, Netflix user ratings, and, famously, Latanya Sweeney’s use of the Group Insurance Commission health insurance database to identify then-Massachusetts Governor William Weld.