The "Re-identification" of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risks and Privacy Protections, Then and Now

Author: Daniel C. Barth-Jones, M.P.H., Ph.D., Assistant Professor of Clinical Epidemiology, Department of Epidemiology, Mailman School of Public Health, Columbia University.

Abstract:

The 1997 re-identification of Massachusetts Governor William Weld's medical data within an insurance data set which had been stripped of direct identifiers has had a profound impact on the development of de-identification provisions within the 2003 Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Weld's re-identification, purportedly achieved through the use of a voter registration list from Cambridge, MA is frequently cited as an example that computer scientists can re-identify individuals within de-identified data with "astonishing ease". However, a careful re-examination of the population demographics in Cambridge indicates that Weld was most likely re-identifiable only because he was a public figure who experienced a highly publicized hospitalization rather than there being any certainty underlying his re-identification using the Cambridge voter data, which had missing data for a large proportion of the population.

The complete story of Weld's re-identification exposes an important systemic barrier to accurate re-identification known as "*the myth of the perfect population register*". Because the logic underlying re-identification depends critically on being able to demonstrate that a person within health data set is the only person in the larger population who has a set of combined characteristics (known as "quasi-identifiers") that could potentially re-identify them, most re-identification attempts face a strong challenge in being able to create a complete and accurate population register. This strong limitation not only underlies the entire set of famous Cambridge re-identification results but also impacts much of the existing re-identification research cited by those making claims of easy re-identification. This paper critically examines the historic Weld re-identification and the dramatic reductions (thousands fold) of re-identification risks for de-identified health data as they have been protected by the HIPAA Privacy Rule provisions for de-identification since 2003. The paper also provides recommendations for enhancements to existing HIPAA de-identification policy, discusses critical advances routinely made in medical science and improvement of our healthcare system using de-identified data, and provides commentary on the vital importance of properly balancing the competing goals of protecting patient privacy and preserving the accuracy of scientific research and statistical analyses conducted with de-identified data.

Keywords: HIPAA, privacy, de-identification, statistical disclosure, population register, quasi-identifiers, K-anonymity, public policy

The "Re-identification" of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risks and Privacy Protections, Then and Now

Author: Daniel C. Barth-Jones, M.P.H., Ph.D., Assistant Professor of Clinical Epidemiology, Department of Epidemiology, Mailman School of Public Health, Columbia University.

Introduction:
Dateline 1996-1997: Collapse and Attack

May 18, 1996, was a cooler-than-usual spring day, but Massachusetts Governor William Weld wasn't feeling well under his cap and commencement gown. Governor Weld was to receive an honorary doctorate degree from Bentley College and give the keynote graduation address. But, unbeknownst to him, he would instead make a critical contribution to the privacy of our health information. As he stepped forward to the podium, it wasn't what Weld said that now protects your health privacy, but rather what he did: He teetered and *collapsed unconscious* before a shocked audience.

Weld's contribution to this saga essentially ended with this episode. But others would step forward in what would become an influential event in U.S. public policy for protecting health privacy. Video of Weld's collapse was repeatedly aired by local television news outlets in what some described as a politically motivated "collapse-a-thon," purportedly intended to degrade the Republican governor. Newspapers dutifully reported that Weld had been taken by ambulance to Deaconess Waltham Hospital, given a stress test and other diagnostic procedures, and then discharged the following day with a diagnosis of influenza – all information that would be routinely recorded as part of the process of paying Weld's insurance bill.

Weld recovered quickly. The months passed and this incident might have passed quietly, but for Latanya Sweeney, then a MIT graduate student in computer science. Sweeney's studies in computational disclosure control had brought to her attention hospital data released to researchers by the Massachusetts Group Insurance Commission (GIC) for the purpose of improving healthcare and controlling costs.

In his 2010 UCLA Law Review paper, "*Broken Promises of Privacy*," *(Ohm, 2010)* University of Colorado law professor Paul Ohm describes Sweeney's re-identification of Weld's hospitalization data:

> *"At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then-graduate student Sweeney started hunting for the Governor's hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office."*

Ohm's version of the story —frequently retold in the blogosphere—is a disturbing tale that should trouble anyone concerned about the privacy of health data. Not surprisingly, this single case of

re-identification has had broad influence on both the development of U.S. health privacy policy and on public opinion about the risks posed by supposedly "de-identified" health data. Fortunately, the privacy protections that have since resulted through the establishment of key provisions in the 2003 HIPAA[1] privacy rules work well to provide very effective—but largely unseen and unsung— privacy and health benefits to all Americans.

Recently though, some voices in the privacy arena - like Ohm - have harkened back to the Weld re-identification, using it as support for their public fretting that computer scientists can purportedly identify individuals hidden in anonymized data with "astonishing ease."

However, the complete story of Weld's re-identification is more complex and exposes a systemic limitation that underlies much of the research cited by those who make claims of startlingly easy re-identification. In fact, a careful examination of the population demographics in Cambridge indicates that Weld was most likely re-identified only because he was a public figure who experienced a highly publicized hospitalization rather than there being any real certainty about the accuracy of his attempted re-identification using the Cambridge voter data. And, although Weld clearly couldn't have been re-identified by a voter list attack today because of the HIPAA privacy protections that the event has helped to inspire, detailed scrutiny of this historic attack is still warranted because the report of this re-identification skims over a crucial fact: the Cambridge voter-roll used in the attack (with 54,805 persons) *(Sweeney, 1997a)* actually included only just more than half of the true population of Cambridge at the time of Weld's collapse. This introduces a "fatal flaw" into the logic used to purport Weld's re-identification using voter registration data. The problem, which I'll call more generally the "*myth of the perfect population register",*[2] is that, without a complete and accurate listing of the Cambridge population, one could not be certain that Weld was actually the only male with his birthday in his ZIP code. The same sort of fatal flaw affects the entire set of associated Cambridge re-identification attack results, *(Sweeney, 1997a)* and, furthermore, has important implications for other research done on the estimating of re-identification risks because similar flaws routinely occur within other such studies.

Because a vast array of healthcare improvements and medical research critically depend on de-identified health information, our essential public policy challenge then is to accurately assess the current state of privacy protections, and balance risks and benefits to maximum effect.

How Typical was Weld's Re-identification?

Admittedly, Governor Weld was easy to re-identify within the GIC hospitalization data for Massachusetts employees. Because Weld was the governor and was publicly known to have been hospitalized, one could expect that Weld's hospital billing data would be within the GIC hospital data set. This foreknowledge would not likely exist for random re-identification targets unknown to an imagined "data intruder" (trade jargon for someone who is out to undertake this sort of re-identification malfeasance). For a randomly selected target, a data intruder would be unlikely to know whether any chance target individual was a state employee or had been recently hospitalized.

In fact, Weld's "shooting-fish-in-a-barrel" re-identification lacked almost all of the challenges that would typically exist for most re-identification attempts. Because Weld was known to reside in Cambridge and likely the subject of ballet casting photo-ops, he was also certain to be listed in the Cambridge voter registration rolls – and so another key hurdle required for a re-identification attempt using a voter list attack was overcome. The list for Cambridge contained data for 54,805

[1] Health Insurance Portability and Accountability Act of 1996.
[2] See Paul Ohm's excellent cautionary tale for public policy-makers in his paper entitled the "*Myth of the Superuser*" (*Ohm, 2008*), which adeptly explains how inappropriate conflation of the rare and anecdotal accomplishments of notorious hackers with the actions of typical users often forms distorted views of the normative behavior under consideration for regulatory control, thus leading to poorly constructed public policy. Important parallels exist here with regard to the ability of most data intruders to construct accurate and complete population registers capable of supporting re-identification attacks.

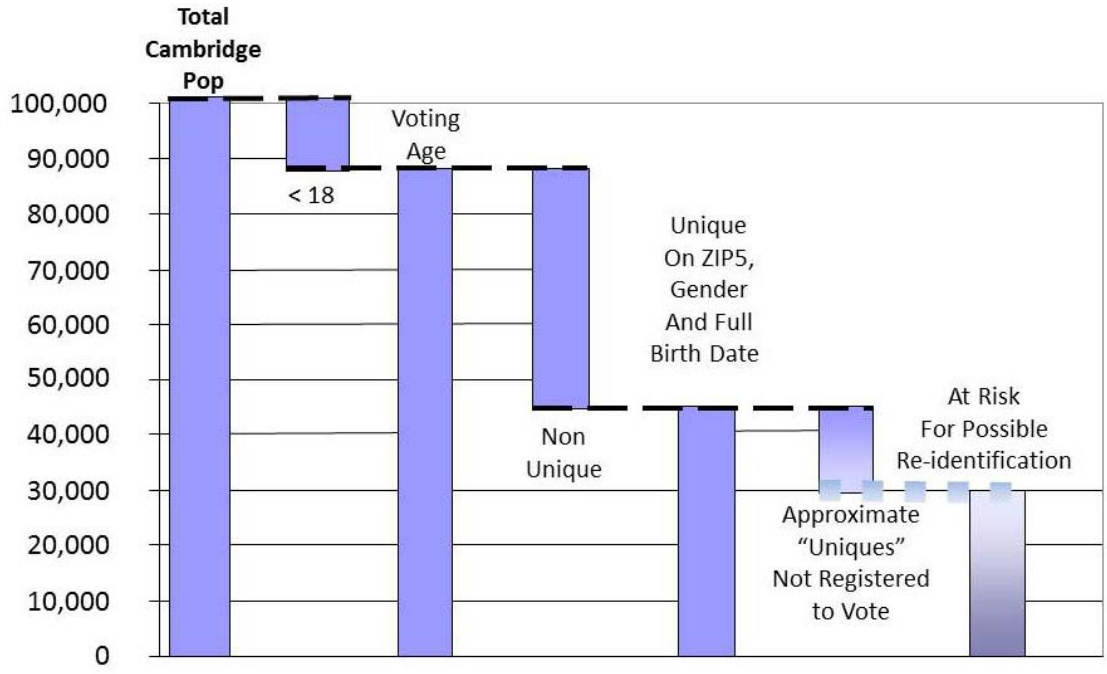
individuals and included their full date of birth, ZIP code and gender. However, as seen in Table 1, the population of Cambridge in 1996-97 was somewhere close to 100,000 persons (the 1990 U.S. Census showed 95,802 residents and the 2000 Census showed 101,391). Despite the high drama surrounding Weld's re-identification, almost half of the Cambridge population could not have possibly been re-identified by such a voter registration list attack.

Table 1. Cambridge, MA Population and Registered Voters at Time of 1996-97 Weld Attack

U.S. Census Population Counts and Estimated 1996-97 Total Population for Cambridge, MA		Percent
Total Cambridge, MA Population in 2000 Census	**101,391**	
Total Cambridge, MA Population around 1996-1997*	**99,435**	**100%**
Total Cambridge, MA Population in 1990 Census	**95,802**	
Individuals in 1997 List Used for Weld Attack	**54,805**	**55%**
Estimated Unlisted Population	**44,630**	**45%**
* Estimated by linear interpolation from 1990 and 2000 Census data		

Here it is useful to further explore the 2000 U.S. Census data to obtain some sense of how typical Weld's re-identification might have been. Using the 2000 Census data for Cambridge, Figure 1 provides a graphic illustration of the proportion of the population that could have been subject to a potential re-identification risk using a voter list attack.

Figure 1. Estimated Proportion of the Cambridge Population Subject to Potential Re-identification Risk



The 2000 Census shows that approximately 13,000 persons in the Cambridge population were under the age of 18 and, therefore, would have been among those who were not in the voter list. The Census data also provides us with a cross-tabulation of age (in years) by gender and five-

digit ZIP code from which we can estimate how likely it is that people would be unique on their combination of full date of birth, gender and five-digit ZIP code in Cambridge. These "Probability 101" *pigeon-hole principle* calculations[3] *(Golle, 2006)* show that about 45,000 persons in Cambridge who were above age 18 were likely to be unique by this combined set of characteristics, or about 51 percent of the 88,000 persons in Cambridge's voting age population. However, some 31,000 (35 percent) of the 88,000 persons in Cambridge's voting age population were not on the list used to attempt re-identification. So if we assume, just for argument's sake, that non-registration wasn't associated with age, gender or ZIP code as a straw-man premise, then we would arrive at a "ballpark" estimate that perhaps 16,000 (~35 percent) of the 45,000 potentially identifiable "population uniques"[4] in Cambridge would be likely to be directly protected from re-identification by virtue of their non-registration.

This leaves roughly 29,000 persons (about 29 percent of Cambridge's entire population) at some plausible risk of re-identification using the combination of date of birth, gender and 5-digit ZIP code—if, and this is a big "if", they had also been in the GIC hospital data set.[5] Producing precise estimates of the number of individuals at potential risk of re-identification would require access to both the GIC hospital data and the 1996-97 age, gender and ZIP code stratified voter list fractions. But having a precise estimate is not important for our arguments that follow here. While a 29 percent risk of *possible* re-identification isn't reassuring, it is essential to press on with our scrutiny here because this is only the risk of *possible* re-identification, and not *actual* re-identification. The true risk of re-identification would likely be *even lower* because about 31,000 (~35 percent) persons in the voting age population were not in the voter list, so many of the individuals who looked unique within the voter list would not actually have been the only person in Cambridge to have their particular set of birth date, gender and ZIP code characteristics.

Figure 2 illustrates that the precise conditions required for definitive re-identification can be quite daunting. Note that this figure illustrates only those limited cases where only one or two persons with shared "quasi-identifier" characteristics (like combined ZIP code, birth date and gender) exist in either the healthcare data set or in the voter registration list. For all of the many other cases where more than two individuals would share the same quasi-identifiers in the healthcare or voter data, certain re-identification is not possible. Rows 1 and 2 in the figure show that an individual must be in both the voter list and the healthcare data set in order to be at risk of re-identification. Row 3 illustrates that, at least without further detailed information in the health data already being known about a target individual, when more than one person exists in the health data with the same quasi-identifier combination, the individual cannot be definitively re-identified. Row 4 illustrates that, when more than one person in the voter list shares the same quasi-identifier set, definitive re-identification is also not possible. Finally, Row 5 illustrates the high likelihood (50 percent) of false re-identification when a "false positive" arises due a person missing from the population register with the same quasi-identifier characteristics. If a single person is missing with the quasi-identifiers in question, then the incorrectly supposed "re-identification" is *no better than a flip of a coin,* and the probability of correct re-identification becomes even lower when there is more than one person missing from the population register.

Figure 2. Re-identification Failure and Success Conditions

[3] Philipe Golle used these same methods to demonstrate that about 63 percent of the U.S. population could be potentially re-identifiable through their combined characteristics for full date of birth, five-digit ZIP code and gender.

[4] Statistical disclosure jargon for individuals who are the only person with their combined set of indirect identifiers.

[5] It seems safe to presume that only a relatively small proportion of Cambridge residents would be both state employees and hospitalized within this time period.

	HOSPITAL DATA SET (Found In Data Set)	VOTER DATA SET (Found in Data Set)	NON-VOTERS (in Population)
1	**Not in Hospital Data**	Male 1/1/1945 02138 **Can't Re-identify (No Match)**	
2	Male 1/2/1945 02138	**Not in Voter Data**	Male 1/2/1945 02138 **Can't Re-identify (No Match)**
3	Male 1/3/1945 02138 Male 1/3/1945 02138	Male 1/3/1945 02138 **Can't Re-identify (> 1 Match)**	
4	Male 1/4/1945 02138 **Can't Re-identify (> 1 Match)**	Male 1/4/1945 02138 Male 1/4/1945 02138	
5	Male 1/5/1945 02138 ***Presumed* Re-identification (Has Only 50% Chance of Being a Correct Match)**	Male 1/5/1945 02138	Male 1/5/1945 02138 **Directly Protected From Re-identification**
6	Male 1/6/1945 02138 ***Correct Re-identification***	Male 1/6/1945 02138	

The annotations in Row 5 also point out that every person not within the voter list is directly protected from being re-identified, and, furthermore, their absence from the population register also reduces the probability that others who share their quasi-identifier set would be correctly re-identified. This is an extremely important limitation on re-identification when imperfect population registers are used. Nationwide, about 29 percent of the voting age population is not registered to vote. *(File, et al., 2010)* These non-voters are directly protected from re-identification attempts using voter registries; but, they also importantly confound the attempts to re-identify those registered to vote whenever such incomplete voter registries are used.

Without the important advantage of the auxiliary information available regarding Weld's public hospitalization, a data intruder would have had to go through the daunting process of making sure that there were not any other males living in the ZIP code 02138 at the time of Weld's collapse who were born on Weld's birthday in order to be certain that Weld was correctly re-identified using such a voter list attack method. Because there were approximately 35,000 persons living in ZIP code 02138 in 1997, it is difficult to imagine how any lone data intruder would have had the ability to complete this essential step in the re-identification process. A realistic assessment of the validation effort involved in trying to accurately track down all of the 50-ish males among 35,000 persons and verify their birth dates would certainly seem a wearisome and time-consuming endeavor at best. Clearly, many more tools exist in today's information rich world than existed

back in 1997. But even with all of today's tools, online data frequently contains errors and some people will always be missing from any easily obtained source of data. The reality facing a would-be data intruder is that in addition to frequent errors in online information, people move (and don't always promptly update their address information); and some segment of any population is simply "off the grid". Such problems as differences in data coding between data sets, real changes in variables occurring over time and plain-and-simple data errors all lead to "data divergence". *(Duncan, et al. 2011)* And as we've seen in Figure 2, every error in our population register protects at least one person[6], with some errors -- like persons missing from our population register -- going even further by providing probabilistic protection against definitive re-identification to other individuals within the population.

However, this tiresome task of attempting to create a perfect population register was not required in the Weld re-identification because of the distinct advantage of being able to know in advance of the re-identification attack that Weld was admitted to Deaconess Waltham Hospital on May 18, 1996, given a stress test, chest x-ray, EKG and other tests during his stay and then discharged the following day with a diagnosis of influenza. Weld's birth date and ZIP code were in the public record as well, and provided an additional check on the match (although the diagnosis, procedures and admission/discharge dates provided quite a reliable means to make a confident match.) A simple confirmation of these well-publicized facts through examination of the hospital data would have been sufficient to gather one's confidence and send Weld his hospital data from the GIC data set without any further need for the exhaustive verification step involved in confirming that there were not any other male residents of ZIP 02138 who shared his birthday. The whole process of linkage to the voters list was superfluous to Weld's re-identification—it was not necessary and, more importantly, it was not sufficient. It may have been reassuring that Weld was the only male voter with his birthday within 02138, but in light of what could have already been known from the extensive publicity surrounding the specifics of Weld's hospitalization, there was not any need for voter list linkage to feel confident that Weld's record in the GIC hospital data had been re-identified.

Let's consider how reliable a re-identification attack could have been under a scenario where the data intruder did not possess the highly publicized additional information that existed in the clearly atypical Weld example. A data intruder without this additional knowledge would not have had any easy means of being able to know with certainty exactly how many 50-year-old, male unregistered voters there were in ZIP 02138 in 1997. However, three years later the 2000 U.S. Census reported that there were a total of 174 50-year-old males in 02138. Assuming for argument's sake that this same number was present in 1997 and following the same pigeonhole probability calculations referenced earlier *(Golle, 2006)* for the expected number of days on which two or more persons will share a birthday in the year, we find that there would have been only a 62 percent chance that Weld was correctly identified as the only 50-year-old male with his same birthday in 02138.

If we examine both the 1990 and the 2000 Census results in order to get some idea of the reasonable bounds for the likely variation for the expected number of 50-year-old males within 02138 in 1996, we find that it would be quite surprising if there were fewer than 150 persons with these combined characteristics.[7]

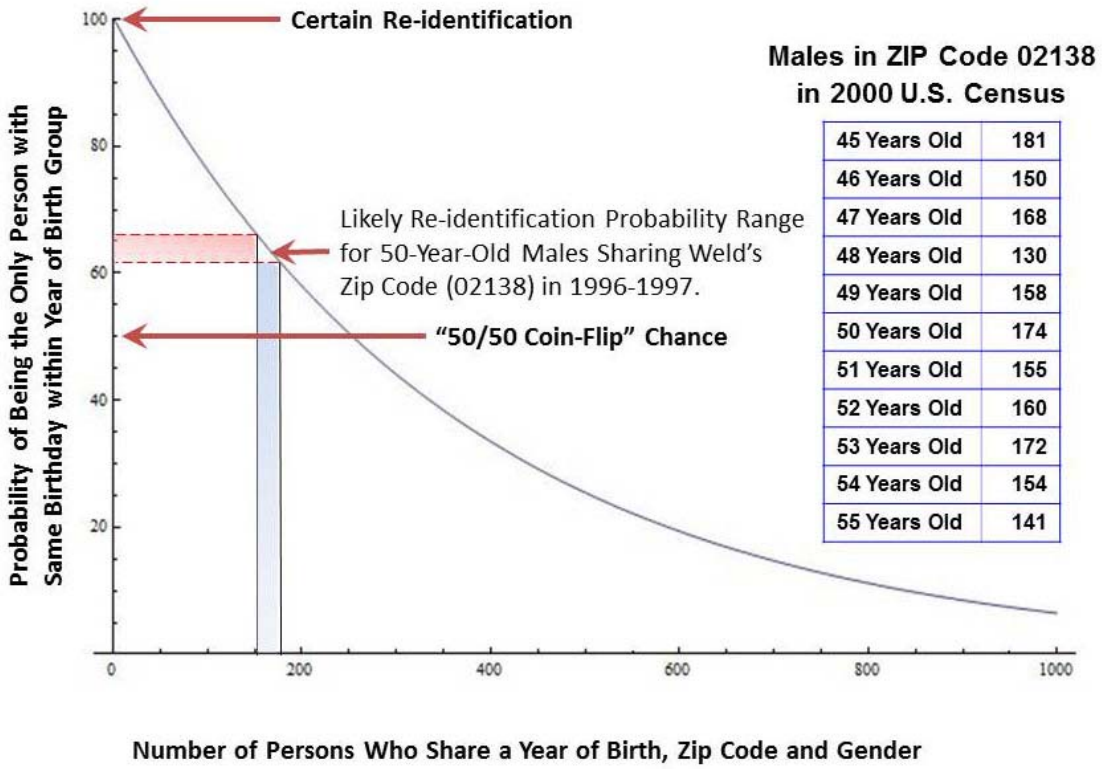
---

[6] To be thorough in our discussion, it should be mentioned that sophisticated data intruders might make use of probabilistic data linkage methods which may overcome some limited types of data errors, but such probabilistic linkage is inherently subject to uncertainty as to whether a definitive re-identification can be made and, furthermore, real world data intruders would rarely, if ever, be in a position to test and verify the extent of this potentially substantial uncertainty.

[7] Thoughtful examination of the available data leads to a number of approaches that could be considered when assessing what can be reasonably assumed about the number of 50-year-old males in Weld's ZIP code at the time of this collapse and the subsequent re-identification attack. For example, should we "age" the 1996-97 population forward three to four years into the year 2000 as if the population was static and had no immigrations, emigrations or deaths? Doing this, persons who were 50 in 1996 would be assumed to be 53 -54 years old in 2000 depending on the particulars of when their birthdays fell relative to the April 1, 2000 Census count date. Alternatively, we could assume a steady-state age distribution and use this observed year-to-year variation in ages to estimate the likelihood of immigration, emigration or

Figure 3 below shows that, of the 150 to 175 50-year-old males residing in 02138, the probability of there being another unregistered voter who shared Weld's exact birthday would have been between 32 and 38 percent, and the assertion that Weld was re-identified on the basis of the voter linkage had–at best—only about a two thirds' chance of being correct.

Figure 3. Probability of Re-identification for 50-Year-Old Males Sharing Weld's Voter Registration Information.



While somewhat better than a flip of a coin, this 62 to 66 percent probability of accurate re-identification yields little confidence that Weld was actually "re-identified" on the basis of the voter linkage attack. There was, after all, about a 35 percent chance that the alleged re-identification was incorrect. Most people reading that Weld was re-identified using voter data are likely to assume that this "re-identification" was made with certainty and had been definitively accomplished via the linkage with voter data. Yet, even if we take Weld's "re-identification" as a probabilistic statement, a 35 percent chance for error greatly exceeds the usual p-value standards of one percent (or even five percent) at which scientists have typically declared an event to be "statistically significant" and to have been unlikely to have occurred by chance.[8] So,

---

death events in this age range. Experimenting with all of these assumptions in order to test the robustness of our arguments, we consistently end up with fairly unwavering findings. From the available U.S. Census data, it seems likely that there were at least 150 to 175 males residing in Weld's 02138 ZIP code and sharing his year of birth cohort in 1996-97, so there would have been about a 33 to 38 percent chance that Weld shared his birthday with at least one other male in his ZIP code.

[8] It is also worth noting that Weld wore a demographic bull's eye within this college-age dominated Cambridge ZIP code located near MIT and Harvard. If one had attempted to re-identify one of the thousand 20-year-old males in Weld's ZIP code, even with a perfect population register, probability calculations show that only about 6 percent of them would have

without benefit of the news coverage regarding Weld's public collapse and hospitalization, the now famous Weld "re-identification" might have never become the touchstone for privacy reform that it has become today.

It's difficult to overstate the influence of the Weld/Cambridge voter list attack on health privacy policy in the United States. In the November 3, 1999, Notice of Proposed Rulemaking (NPRM) for the HIPAA Privacy Rule published in the Federal Register, Department of Health and Human Services (HHS) policy-makers state that:

> "*A 1997 MIT study showed that, because of the public availability of the Cambridge, Massachusetts voting list, 97 percent of the individuals in Cambridge whose data appeared in a data base which contained only their nine digit ZIP code and birth date could be identified with certainty." (Federal Register, 1999, citing Sweeney, 1997b)*

It is not particularly clear from HHS's discussion in the NPRM preamble as to what their full understanding was of the many complexities and methodological flaws raised here regarding the purported re-identification estimates for the Cambridge study, but the influence of this work on the development of the de-identification rules in the HIPAA Privacy Rule is unquestionable. In response to concerns about the requirements for "Safe Harbor" de-identification from the Proposed Rule appearing in the December 28, 2000 Federal Register, HHS responded:

> "*...we remain convinced by the evidence found in the MIT study that we referred to in the preamble to the proposed rule and the analyses discussed below that there remains a significant risk of identification of the subjects of health information from the inclusion of indirect identifiers such as birth date and ZIP code..." (Federal Register, 2000)*

With the benefit of hindsight, it is apparent that the Weld re-identification has served an important illustration of privacy risks that were not adequately controlled prior to the advent of the HIPAA Privacy Rule in 2003. It is now quite clear that simple combinations of high resolution variables, like birthdates and ZIP codes, can put an unacceptable portion of the population at risk for potential re-identification. So even though Weld's "re-identification" plainly wasn't achievable using the now-famous Cambridge voter list linkage attack, his unfortunate collapse in 1996 has still been a positive force for American privacy policy. The message learned from the Weld re-identification is now widely understood: *We **must** pay attention to the potential for re-identification when data has only been stripped of the directly identifying information such as names and addresses*.

The Myth of the Perfect Population Register

However, what must also be unmistakably understood by policymakers within the privacy arena is a much broader issue posed by this example: A somewhat furtive "insider" trade secret underlies most similar work conducted by disclosure risk scientists.

The same "perfect population register" Achilles heel that limited the accuracy of the Weld re-identification underlies many, if not most, of the re-identification risk estimates made by statistical disclosure risk scientists. Creating a "perfect population register" that is complete and accurate is a tremendous challenge for even the U.S. Census Bureau and would typically be far beyond the likely abilities of a hypothetical data intruder. Disclosure risk scientists themselves usually cannot afford to complete this final exhaustive step when making re-identification risk estimates. So they wisely skip this last essential task and instead make easily obtained, but highly conservative, estimates of the true re-identification risks. This is an appropriate practice as long as everyone who interprets the results understands that we've left out the hardest part of the equation and

---

had a unique birth date in this large college age population; so over 93 percent of 20-year-old males in Weld's ZIP code would not have been re-identifiable.

chosen to err strongly on the side of caution in order to protect privacy. It has to be recognized though that missing and incorrect data will inevitably plague any attempt to build a perfect population register and, thus, to the extent the population register is imperfect, significant proportions of purported "re-identification" matches may simply be incorrect.

It seems reasonable to presume that data intruders might be able to create near-perfect population registries for small or isolated populations for limited time periods aided by their personal knowledge of the population within a specific location. *(Elliot, et al., 1999)* But because the final step in the re-identification process always depends critically on being able to rule out that there are not individuals missing from the population register and that the quasi-identifier information was correct in both the data source and the population register, every certain re-identification faces the some dauntingly effort-intensive and often very expensive prerequisites. Even with the expansion of available information resources to aid this task, any realistic assessment of a lone data intruder's ability to accurately create population registries which include time-dynamic quasi-identifiers (such as patient locations) for populations numbering in the tens of thousands should include some healthy skepticism about the purported "re-identifications". And, as will be shown in the following discussion of re-identification risks under today's HIPAA protections, just like attempting to confirm that there are "no black swans", it is now a very tall order indeed to verify supposed re-identifications with anything approaching certainty.

Re-identification Risks Today Under the HIPAA Privacy Rule

Fortunately, HHS appropriately responded to the concerns raised by the Weld/Cambridge voter list privacy attack and, under HIPAA Privacy Rules effective in April, 2003, acted to help prevent re-identification attempts by future data intruders. The HIPAA Safe Harbor provision now restricts disclosure of an individual's full date of birth (only year may be reported), caps such reporting at 90 years of age, and restricts the reporting of five-digit ZIP code (with three-digit ZIP codes for populations of greater than 20,000 persons being the smallest reportable geographic unit).

Consider how the same Weld voter list attack would play out under today's HIPAA protections. A data intruder would now only be capable of linking all 50-year old males within the three-digit ZIP code "021" between the hospital data and the voters list. Even if Weld alone held these combined characteristics in the hospital data (considerably less likely to be the case now given the required HIPAA changes), he would be among 7,299 other persons sharing Weld's characteristics in the three-digit ZIP "021". And, even if a data intruder could undertake the overwhelming efforts needed to determine which of 7,299 50-year-old males in three-digit ZIP "021" linked to such a unique hospital record, he would still have a population of 1.25 million in which it would be necessary to confirm that there was not at least one other 50-year-old male who had not registered to vote in order to confidently claim a re-identification.

Broadening further to examine the risks for the entire Cambridge population, if we assume that a data intruder could somehow construct a perfect population register for the 1.25 million persons living in this three-digit ZIP code "021" which includes much of the Boston area, no unique combinations of birth year, gender, and ZIP code “021" exist in either the 2000 or 2010 Census data when the Safe Harbor de-identification standard is used.[9] Under present HIPAA rules, this re-identification risk has been reduced to the point where we can no longer reliably detect it.

This particular post-HIPAA re-identification scenario makes a useful illustration of the Weld/Cambridge case; but, to avoid the all-too-frequent trap of arguing policy positions only from carefully selected examples, we turn to the August 23, 2007, testimony before the *Ad Hoc Workgroup on Secondary Uses of Health Data* of the *National Committee on Vital and Health*

[9] Although disclosure control methods were used in the reporting of the U.S. Census data, the nature of the data-swapping and reporting controls used are such that it is clear unique individuals would, at most, only very rarely exist within the 3-digit ZIP code "021", so the re-identification risk for these combined characteristics is extremely small.

Statistics. Dr. Sweeney, by then an accomplished professor of computer science at Carnegie Mellon University, reported that 0.04 percent (4 in 10,000) of the individuals in the U.S. population within data sets de-identified using the Safe Harbor method could be identified on the basis of their year of birth, gender and three-digit ZIP code.*(NCVHS, 2007)* While the details of the methods underlying this Safe Harbor re-identification estimate also warrant further attention, useful public policy insights can be gained by taking these Safe Harbor estimates at face value. Given that voter lists are far from perfect population registries, we would estimate that, of the one in 2,500 persons at *possible* risk of re-identification from a voter registration re-identification attack, approximately one third of those would likely directly escape re-identification because they are not in voter data.[10] This means the risk of re-identification for Safe Harbor data using a voter list attack would be approximately one in 3,500 - if someone would bother to attempt re-identification with such a miniscule chance of success and considerable additional uncertainty as to whether the re-identification was actually correct as opposed to being a "false positive".

To put some perspective on this estimate, according to National Weather Service statistics, consider that this risk falls somewhere between one's lifetime odds of being personally struck by lightning (about one in 10,000) and the risk of being affected because someone close to you has been struck (with ten people affected for every one struck).*(NWS, 2012)* Note that even if we assume the data intruder could incur the inordinate costs of purchasing a more complete population register for the U.S. from commercial sources rather than using voter registration data, this one-in-2,500 Safe Harbor possible re-identification risk estimate continues to fall in this "*should I really worry about me or my loved ones being hit by lightning?*" range of risks.

Further boosting our confidence that re-identification is no longer a trivial task for anyone with 20 dollars and malicious intentions are two recent studies by disclosure risk scientists. Kathleen Benitez and Bradley Malin's 2010 study estimated re-identification risks under the HIPAA Safe Harbor rule on a state-by-state basis using voter registration data, evaluating the expense involved in accomplishing a potential re-identification.*(Benitez et al., 2010)* The percentage of a state's population estimated to be vulnerable (i.e., meaning *not* definitively re-identified, but *potentially* re-identified using the typical simplifying assumptions of a hypothetical perfect population register) ranged from 0.01 percent to 0.25 percent, based on U.S. Census estimates. Additionally, when Benitez and Malin considered state-specific data reporting, coding, and data costs for voter list attacks, the risks further dropped for many states, and for some states the risk was zero due to the lack of key information in their voter registries. They also found that costs per person re-identified ranged from $0 to $17,000, confirming considerable variability in the feasibility of voter registration based attacks.

In an attempt to elucidate what would realistically be involved in an actual re-identification attack, the HHS Office of the National Coordinator for Health Information Technology (ONC) conducted a case study in 2011 examining an attack on HIPAA de-identified data under realistic conditions. The ONC put together a team of statistical experts to assess whether data properly de-identified under HIPAA could be combined with external data to re-identify patients. The study was performed under practical and plausible conditions and *verified the re-identifications against direct identifiers—a crucial step often missing from this sort of study*. The team began with a set of about 15,000 patient records that had been de-identified in accordance with HIPAA and studied the hospital admission records of Hispanics in one hospital system between 2004 and 2009. The data set was stripped of identifying information as required by the HIPAA Safe Harbor methodology. The ONC simulated an intrusion scenario in which an intruder paid for access to an expensive commercial data source rather than using inexpensive voter data and used this data in its attempt to identify specific people from with the HIPAA de-identified data. To verify the study results, the hospital system supplying the data independently confirmed the number of the suspected matches which were real. The experiment showed a match for only two of the fifteen

[10] Other more complete commercial data sources might possibly be used, but accounting for the considerable cost and effort involved in such an attack and the associated accuracy of re-identification is worthy of its own separate examination at another time.

thousand individuals (a re-identification rate of 0.013 percent), and even when maximally conservative assumptions were made about the possible knowledge of the hypothetical intruder, the re-identification risk (under the questionable assumption that re-identification would even be attempted) was likely to be less than 0.22 percent. *(Kwok et al., 2011)*

Clearly, under the existing HIPAA de-identification requirements, re-identification has become expensive and time-consuming to conduct, requires serious computer and mathematical skills, is rarely successful, and almost always ultimately uncertain as to whether it has actually succeeded. The inability of seat belts, airbags and door locks to definitively protect us from harm in car crashes and burglaries does not lead us to the conclusion that we should abandon their use. Likewise, a prudent public policy perspective would not consider abandoning the dramatic protections provided by today's HIPAA de-identification provisions simply because they might rarely fail.

What Have We Learned, and What's at Stake for the Future of Health Care?

Public policy tends to progress in a series of pendulum swings. Two steps forward, one step back. Societies often ignore important and obvious risks for far too long until the unfortunate event in question occurs. Then, legislators and bureaucrats step in with a plan to address the unattended problem and its consequences. Next, the corrective action has its own unintended ripple effects that extend far beyond the initial problem. So again, we fine-tune our rules and regulations. By the time this cycle has completed, the world has often changed, and new circumstances require that we begin again. And on it goes.

Re-identification risks under the current HIPAA Privacy Rule have been reduced to the point that most people wouldn't (and shouldn't) lose any sleep over the issue. Nevertheless, the Weld re-identification story is still frequently reported in an echo chamber of blogs and news reports as if it reflected the current realities under the existing HIPAA protections. For a 15-year-old story predating the advent of HIPAA, the legend of the voter list "re-identification" of Weld seems to have left the realm of useful public policy motivator and to have now entered the realm of "urban myth."

Balancing Privacy Protection and Scientific Accuracy

Unfortunately, considerable costs come with incorrectly evaluating the true risks of re-identification under current HIPAA protections. It is essential to understand that de-identification comes at a cost to the scientific accuracy and quality of the healthcare decisions that will be made based on research using de-identified data. Balancing disclosure risks and statistical accuracy is crucial because some popular de-identification methods, such as "k-anonymity methods," can unnecessarily, and often undetectably, degrade the accuracy of de-identified data for multivariate statistical analyses. K-anonymity requires that original data be changed until multiple persons have been made to share common values for each combination of quasi-identifiers, which distorts the data in ways well understood by statisticians and computer scientists, *(Aggarwal, 2005) (Brickell, et al., 2008)* but sometimes not well-appreciated in the public policy arena. Poorly conducted de-identification and the overuse of de-identification methods in cases where they do not produce real privacy protections can quickly lead to "*bad science*" and damaging policy decisions. These critical considerations are, regrettably, greatly underappreciated by some involved in the ongoing health privacy policy debate surrounding de-identification.

When we over de-identify data, we destroy crucial statistical relationships and correlations within the data with no benefit because of the unavoidable and inherent trade-off between re-identification protection and information quality. Even worse, if we abandon the use of de-identified data because we falsely believe that de-identification cannot provide valuable privacy protections, we will lose the rich benefits that come from analysis of de-identified health data. Jane Yakowitz, a University of Arizona Law School Professor, wrote extensively on this topic in her paper, "Tragedy of the Data Commons," and addresses the societal costs in information flow and knowledge growth that would follow the abandonment of a realistic assessment of the risks of re-identification.*(Yakowitz, 2011)* The reality is that, while one can point to very few, if any, cases of persons who have been harmed by attacks with verified re-identifications,*(El Eman, et al., 2011a)* virtually every member of our society has routinely benefited from the use of de-identified health information. De-identified health data is the workhorse that routinely supports numerous healthcare improvements and a wide variety of medical research activities.*(El Eman, et al., 2011b)* But in the same way that it would be difficult to point to exactly who has had their lives saved by speed limit laws, it should be quite clear that some among us owe our lives to the ongoing research and health system improvements that have been realized because of the analysis of de-identified data. Hopefully, these advancements will continue in generations to come, but unfounded fears of re-identification could have the power to derail this progress.

In my own career as an HIV epidemiologist, I have heightened concerns not only for the very important personal privacy of individuals, but also for the serious tragedies that would occur if fears about de-identification led to a failure to detect and control the next emerging infectious disease that begins to spread globally. If we abandon the use of de-identified data simply because of unwarranted fears regarding privacy risks under today's HIPAA regulations, the consequences of such misguided public policy could be truly disastrous. Privacy advocates and policy makers alike must better understand that, rather than posing new privacy risks, using de-identified data under HIPAA results in vast (thousands-fold) improvements in our individual privacy protection and also sustains a rich public good in research and healthcare improvements. This critical role that de-identified health information plays in improving healthcare is becoming increasingly more widely recognized,*(McGraw, et al., 2009),(Peddicord et al., 2010)* but properly balancing the competing goals of protecting patient privacy while also preserving the accuracy of research requires policy makers to realistically assess both sides of this coin. *(McGraw, 2012)* De-identification policy must achieve an ethical equipoise between potential privacy harms and the very real benefits that result from the advancement of science and healthcare improvements which are accomplished with de-identified data. Properly implemented de-identification complying with the HIPAA de-identification provisions goes a long way toward promoting such a reasonable balance, but I would suggest that there is still room for further improvements in this regard.

Where Should We Go From Here?

Perhaps the last lesson from the Weld saga is that, because re-identification attacks could still put rare but very real people—with names, faces, and personal lives—at risk of potential privacy harms, we should actively prohibit re-identification, and require those with access to de-identified data to guard and use it appropriately.

HHS Office of Civil Rights (OCR) regulators have promised to provide new guidance in the near future for the de-identification of health data in response to a Congressional mandate to do so. HHS OCR regulators should consider whether it is appropriate for de-identified data to fall entirely outside of the purview of the Privacy Rule, or whether, like the so called "Limited Data Sets" (LDSs), which have been stripped of 16 types of direct identifiers, de-identified data should be subject to certain terms in required Data Use Agreements (DUAs) or subject to direct HHS mandates for use conditions. Effective parallels to the LDS DUA can be carefully constructed to provide assurances which help to further limit re-identification concerns, but which also impose little unnecessary burden on appropriate uses of de-identified data. Several recommended best

practices for the use of de-identified data which should be considered by regulators as possible mandatory de-identified data use conditions include:

1) Prohibiting of the re-identification, or attempted re-identification, of individuals and their relatives, family or household members.

   HHS should establish civil and criminal penalties for any unauthorized re-identification of de-identified data (and for limited data sets). A carefully designed prohibition on re-identification attempts could still allow Institutional Review Board (IRB) approved re-identification research to be conducted, but would ban any re-identification attempts conducted without essential human subjects research protections.

2) Requiring parties who wish to link new data elements (which might add quasi-identifiers and thus increase re-identification risks) with data de-identified under the "*Statistical De-identification*" provision of the Privacy Rule[11] to confirm that the data remains de-identified consistent with all of the conditions imposed by this provision.

3) Specifying that HIPAA de-identification status would expire if, at any time, the data contains data elements specified within an evolving Safe Harbor list. The Safe Harbor list should be periodically updated by HHS to include new quasi-identifiers for which population registries of sufficient completeness and accuracy might be reasonably constructed. As is the case under the current HIPAA de-identification provisions, Safe Harbor data elements prohibited would be allowed in data sets only if they have been confirmed to have "very small" re-identification risks through the use of the Statistical De-identification provision.

   Of course, the world is not static, and, therefore, the conditions that support re-identification must be expected to evolve and change with advances in technology and data availability. So, in addition to introducing an expiration on de-identification status to account for new Safe Harbor exclusion elements, the statistical de-identification provision might also be updated to indicate that: a) statistical determinations should expire after an appropriate period during which the population distributions of the relevant quasi-identifiers could be expected to have possibly changed importantly,[12] and b) should be required to undergo annual periodic reviews to confirm that there have not been: i) any substantive changes in the external data environment, ii) newly available technologies or methods that support re-identification, iii) changes in regulatory definition of de-identification, or iv) changes in the data elements contained within the de-identified data which would substantially alter the statistical de-identification determination of "very small" re-identification risks.

4) Formally specifying that for statistically de-identified data, anticipated data recipients must always comply with any time limits, data use restrictions, qualifications or conditions set forth in the statistical de-identification determination associated with the data.

5) Requiring that those holding and using de-identified data implement and maintain appropriate data security and privacy policies, procedures and associated physical, technical and administrative safeguards as needed to assure that this data is: (a) accessed only by personnel or parties who have agreed to abide by the foregoing conditions, and (b) will remain de-identified in accordance with HIPAA de-identification provisions.

---

[11] §164.514(b)(1). The statistical de-identification provision in the HIPAA Privacy Rule specifies that health information is not individually identifiable if: “A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable: (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and (ii) Documents the methods and results of the analysis that justify such determination;”.

[12] Three to five years would seem to be an appropriate suggested period based on U.S. demographic changes.

Extensive safeguards might be unnecessary for data with extremely small likelihoods of re-identification. While for other de-identified datasets, such safeguards would seek to assure that the anticipated recipients for statistically de-identified data could reliably be defined and expected, and to ensure that any risks of re-identification attempts are well controlled.

6) Requiring those transferring de-identified data to third parties to enter into data use agreements which would oblige those receiving the data to also hold to these foregoing conditions, thus maintaining an important “chain-of-trust” data stewardship principal accompanying de-identified data throughout its uses.*(Bloomrosen, et al., 2008)*

Data use requirements of the sort suggested above would impose only modest impositions on the use of de-identified data and would help to provide recourse for actions against data intruders and parties who have not properly managed those very small re-identification risks that might still be associated with de-identified data.

Some privacy advocates, like Ohm, have written disparagingly that "*A re-identification ban is sure to fail,.. because it is impossible to enforce."* *(Ohm, 2010)* However, when properly conducted, de-identification should consistently assure that no more than a very small proportion of the de-identified data would be re-identifiable. Thus, economic evaluations using cost-benefit analyses of re-identification attempts reveal that they are not economically viable as small-scale efforts targeted at an individual or small number of individuals, mostly because other more pedestrian methods of "snooping" (e.g. a “nosy neighbor” trash-picking your prescription bottles) would yield much more certain confirmations for dramatically lower costs. To achieve any reasonable economic payoffs sufficient to counter the expense (in terms of time, effort, requisite computer and mathematical skills and data costs) involved in constructing a near-perfect population register, re-identification attempts would need to be large-scale efforts. This will be the case even if we are considering high pay-off attacks like medical identity theft, or attempting to market $100,000 plus per-year prescriptions and biotechnology treatments within our list of re-identification scenarios. Admittedly, marketers pitching treatments with this sort of high pay-off could be expected to tolerate a relatively low certainty of correct re-identifications, but for the fact that each act of marketing outreach would pose yet another opportunity to be caught having attempted re-identification. Because of this, any large-scale re-identification efforts (even when they have high error rates due to uncertain re-identifications) would be likely to be vulnerable to detection using a combination of statistical analyses (similar to those used in investigating employment discrimination allegations) and computer forensics. While it is unlikely that we would be able to catch everyone who attempts re-identification, a few well-publicized prosecutions of data intruders (or even investigations of alleged re-identifiers) would seem likely to deter many others in the marketing arena who might misguidedly contemplate such prohibited actions even when their probability of success is already so small.

Hopefully, HHS regulators will correctly recognize when issuing their impending de-identification guidance that substantive protections for de-identification have already been importantly achieved and will carefully balance the substantial societal benefits that result from our ability to conduct analyses using de-identified health data with any forthcoming de-identification regulations.

Conclusion

Careful examination of the demographics in Cambridge, MA at the time of the re-identification attempt indicates that Weld was most likely re-identifiable simply because he was a public figure who experienced a highly publicized hospitalization rather than there being any real certainty about the accuracy of his attempted re-identification using the Cambridge voter data. In spite of the near legendary status now held by the saga surrounding the linkage of Weld's data to the Cambridge voters list, this act was superfluous to Weld's re-identification and did not provide sufficient evidence to allege any definitive re-identification. Furthermore, the same methodological flaw that undermines the certainty of the Weld re-identification creates widespread and systemic challenges for all data intruders attempting re-identification which should be understood by public

policy-makers seeking to realistically assess current privacy risks posed by HIPAA de-identified data. Construction of perfect (or near-perfect) population registries is a significant, and too often underappreciated, barrier for many common re-identification scenarios.

What are our essential "lessons learned" from this critical examination of the seminal Weld re-identification attack?

I would propose the following:

- Did Weld's "re-identification" have a useful and important impact on improving healthcare privacy? *Yes, quite clearly.*
- Has this event led to important regulations that help to importantly protect patients from re-identification risk? *Undoubtedly.*
- Does the Weld saga realistically reflect the privacy risks that exist under the HIPAA Privacy rules today? *No, not at all.*
- Should we let the minimal risks of re-identification under today's HIPAA protections cause us to abandon our use of de-identified data to save lives and improve our healthcare system? *No, not unless we wish to abandon science and progress in favor of irrational fears.*

Appendix Material: *News releases from May 1996*

Governor Leaves Hospital After Collapse On Saturday
Orlando Sentinel
Dateline May 20, 1996

Gov. William Weld was released from a hospital Sunday after tests showed it was a touch of the flu that caused him to collapse during a college commencement. "I'm not back to 100 percent," Weld said outside Deaconess Waltham Hospital before returning home, "but night and day compared to yesterday." Weld was stricken Saturday while receiving an honorary law degree from Bentley College. He wavered, grabbed at the podium and collapsed into the arms of others on the podium.

Massachusetts Governor Back At Work After Test
Orlando Sentinel
Dateline: May 21, 1996

Massachusetts Gov. William Weld, the Republican U.S. Senate candidate, returned to his Beacon Hill office Monday afternoon after undergoing a stress test. Weld, 50, collapsed Saturday on stage at a college commencement ceremony and was rushed to a hospital. After a full day of tests in the intensive-care unit, doctors decided he was suffering from influenza. He was released on Sunday.

Massachusetts Governor William Weld Collapses During Commencement
By Martin Finucane
AP (as run in *Seattle Times*)

WALTHAM, Mass. - Massachusetts Gov. William Weld collapsed yesterday during commencement at Bentley College, but doctors said they found nothing seriously wrong with him. The 50-year-old governor had just received an honorary doctorate of law when he fainted. "He fell headfirst (toward the podium), but they caught him," said Bill Petras, a graduating senior who sat five rows back from the stage. Weld was briefly unconscious, but was alert by the time he was lifted onto a stretcher and taken to an ambulance. The crowd applauded and Weld waved. Moments before fainting, Weld had started shaking as he approached the podium, Petras said.

Weld, a Republican who is challenging U.S. Sen. John Kerry for his Senate seat in November, had been scheduled to give the keynote address at Bentley's undergraduate commencement, but never got a chance to speak. "Right now, it looks like maybe the flu," said Pam Jonah, one of Weld's press aides, adding that he would stay in Deaconess-Waltham Hospital for 24 hours of observation. Doctors said they performed an electrocardiogram, a chest X-ray and blood tests, but found no immediate cause for concern.

By late afternoon, Weld was in good spirits and asking for reading material, including books by Ernest Hemingway and F. Scott Fitzgerald, Jonah said. Jonah said Weld's wife, Susan Roosevelt Weld, a Chinese-law scholar at Harvard University, was traveling in China. The governor did not call her immediately because he did not want to alarm her unnecessarily, Jonah said.

References:

Aggarwal, C. *On k-Anonymity and the Curse of Dimensionality.* Proceedings of the 31st Very Large DataBase (VLDB) Conference, Trondheim, Norway, 2005:901-9. http://www.vldb2005.org/program/paper/fri/p901-aggarwal.pdf

Benitez, K.; Malin, B. *Evaluating re-identification risks with respect to the HIPAA privacy rule*. Journal of the American Medical Informatics Association. 2010, Vol 17, p. 169-177.

Bloomrosen, M.; Detmer, D. White Paper: Advancing the Framework: Use of Health Data—A Report of a Working Conference of the American Medical Informatics Association. Journal of the American Medical Informatics Association. 2008, Vol. 15(6) p. 715-722.

Brickell, J.; Shmatikov, V. *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data,* ACM International Conference on Knowledge Discovery and Data Mining (KDD08), August 24–27, 2008, Las Vegas, Nevada, USA.

Duncan, G.; Elliot, M.; Salazar-González, J. *Statistical Confidentiality: Principles and Practice.* Springer. 2011, p. 39-40.

El Emam, K.; Jonker, E.; Arbuckle, L.; Malin, B. A systematic review of re-identification attacks on health data. PLoS One 2011; Vol 6(12):e28071.

El Emam, K. ; Jonker, E.; Fineberg, A. *The Case for De-identifying Personal Health Information*. Electronic Health Information Laboratory, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Canada, 2011. Available at SSRN: http://ssrn.com/abstract=1744038 or http://dx.doi.org/10.2139/ssrn.1744038.

Elliot, M.; Dale, A. Scenarios of attack: the data intruder's perspective on statistical disclosure risk. In Netherlands Official Statistics, Volume 14, Spring 1999, Special issue: Statistical disclosure control, p. 6-10.

Golle, P. *Revisiting the Uniqueness of Simple Demographics in the US Population.* Proceedings of the 5th ACM Workshop on Privacy in Electronic Society, 2006, p. 77-80.

File, T.; Crissy S. *Voting and Registration in the Election of November 2008: Population Characteristics*. U.S. Census Current Population Reports. Issued May 2010. http://www.census.gov/prod/2010pubs/p20-562.pdf

Federal Register 45 CFR Parts 160 Through 164 Standards for Privacy of Individually Identifiable Health Information; Proposed Rule, November 3, 1999. p. 59935.

Federal Register 45 CFR Parts 160 Through 164 Standards for Privacy of Individually Identifiable Health Information; Proposed Rule, December 28, 2000. p. 82710.

Kwok, P.K.; Lafky,D. *Harder Than You Think: A Case Study of Re-Identification Risk of HIPAA-Compliant Records*. Joint Statistical Meetings. Section on Government Statistics. Miami, FL Aug, 2, 2011. p.3826-3833.

McGraw, D.; Dempsey, J .X.; Harris, L.; Goldman, J. *Privacy As An Enabler, Not An Impediment: Building Trust Into Health Information Exchange*. Health Affairs Mar/Apr 2009 28:2416-427.

McGraw, D. *Building Public Trust in Uses of Health Insurance Portability and Accountability Act De-identified Data*. Journal of the American Medical Informatics Association 2012. doi:10.1136/amiajnl-2012-000936.

National Committee on Vital and Health Statistics Report to the Secretary of the U.S. Department of Health and Human Services. *Enhanced Protections for Uses of Health Data: A Stewardship Framework for 'Secondary Uses' of Electronically Collected and Transmitted Health Data*, December 19, 2007. Http://www.ncvhs.hhs.gov/ 071221lt.pdf.

National Weather Service. Lightening Safety: Medical Aspects of Lightening. Http://www.lightningsafety.noaa.gov/medical.htm (last accessed June 18, 2012).

Ohm, P*. The Myth of the Superuser: Fear, Risk, and Harm Online.* U of Colorado Law Legal Studies Research Paper No. 07-14; UC Davis Law Review Vol. 41, No. 4, Page 1327, April 2008. Available at SSRN: http://ssrn.com/abstract=967372.

Ohm, P*. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization* (August 13, 2009). UCLA Law Review, Vol. 57, p. 1701, 2010; U of Colorado Law Legal Studies Research Paper No. 9-12. Available at SSRN: http://ssrn.com/abstract=1450006.

Peddicord, D.; Waldo, A.B.; Boutin, M.; Grande, T.; Gutierrez L. Jr. *A Proposal To Protect Privacy Of Health Information While Accelerating Comparative Effectiveness Research.* Health Affairs. Nov 2010 vol. 29, no. 11, p. 2082-2090.

Sweeney, L. *Weaving Technology and Policy Together to Maintain Confidentiality*. Journal of Law, Medicine and Ethics, Vol. 25 1997, p. 98-110.

Sweeney, L. *Guaranteeing Anonymity when Sharing Medical Data, the Datafly System*. Masys, D., Ed. Proceedings, American Medical Informatics Association, Nashville, TN: Hanley & Belfus, Inc., 1997, p. 51–55.

Yakowitz, J., Tragedy of the Data Commons (March 18, 2011). Harvard Journal of Law and Technology, Vol. 25, 2011. Available at SSRN: http://ssrn.com/abstract=1789749 or http://dx.doi.org/10.2139/ssrn.1789749.