Question 1:

One concern regarding big data is the possible compromise of personal information. This has been seen by data breaches of large financial companies resulting in stolen sensitive information such as customer names and credit card numbers. Is “data theft” a legitimate concern with big data analytics? If so, what can be done to protect such a large volume of protected patient information?

Response from Dr. Macedonia:

Information security should be a concern of everyone who handles sensitive data. Virtually everyone who practices medicine in the United States is now part of a vast medical informatics web accessing large data repositories such as their hospital or practice electronic medical record (EMR). The Affordable Care Act (ACA) and the Health Insurance Portability and Accountability Act (HIPAA) have multiple provisions regarding information security. With or without big data analytics, there are vulnerabilities to the security of these systems and everyone who participates in medical care involving sensitive medical data is required by law to secure medical information. Health systems and research institutions are required by law to secure sensitive information as well.

You mention data breaches to financial companies and the theft of credit card numbers, but financial institutions have not been the only data aggregators to be hacked by foreign agents and criminal syndicates. Tens of millions of medical records have been compromised in attacks on two very large health informatics networks, one run by a large health system in the District of Columbia as well as a national for-profit hospital system. These data repositories were not “big data analytics” clusters but rather EMRs maintained for clinical and billing purposes or “small data” applications, if you will. In other words, “big data analytics” is not the major information security problem, but it is intimately connected to vulnerable systems such as EMRs. These are under constant threat and require continuous security measures implemented by all health care institutions and all health research enterprises. Like any entity that exists in this framework, researchers using big data analytics methods should adhere to strict information security policies. Even if there were no legal imperative there is a practical one. To quote from our article, “Just as we place faith in a banking system...”
that utilizes strong accounting principles, prevents money laundering, and provides privacy to the clients and transparency to regulators, we are more likely to use a similar data banking system that adheres to a strict code of standards and practices.”

Question 2:

Are there ethical considerations regarding the collection and dissemination of information from neonates and other minors?

Response from Dr. Macedonia:

Neonates and other minors are generally treated in research ethics under the category of “special populations,” as first outlined in the Belmont Report in the 1970s. This important change to our approach to research in these individuals has, however, had the unintended effect of making research into new treatments for deadly diseases more difficult and more costly to conduct. For this reason, our pediatric colleagues lobbied for and succeeded in passing the Best Pharmaceuticals for Children Act (BPCA) and the Pediatric Research Equity Act (PREA). This legislation has allowed them greater latitude in using observational research and adaptive studies to prove efficacy in drug intervention trials. As a result, pediatric interventions are receiving faster approval with fewer study participants exposed to a lower burden of research-related risk. Our article points to this as a reason why there is an ethical imperative to do the same in the interest of our populations.

Question 3:

How do big data analytics protect against data mining or simply looking through a large volume of information simply to identify associations that may or may not be real?

Response from Dr. Macedonia:

“Data mining” is an older term for big data analytics that is now rarely used in the field because over time it became synonymous with “data dredging” or digging up spurious correlation. It is in nobody’s interest to find spurious correlations and report them as causal connections. Much of the “analytics” part of big data analytics is dedicated to finding true correlation and minimizing spurious correlation. We introduce the reader to some of those existing tools. There is no single method that reliably eliminates this problem 100% of the time, just as there are no reliable methods of eliminating surgical site infections 100% of the time. There are, however, a plethora of mathematical and computational tools available to researchers that can mitigate these risks, and if applied as part of an organized system of data collection and analysis, these tools can validate correlation results.

Question 4:

How can I, as a reader of an article, evaluate the validity of the research methodology reported from big data analytic studies?

Response from Dr. Macedonia:

When I was young I owned cars that I could (and did) disassemble and reassemble in my garage. There came a point when more advanced electronic systems were integrated into cars and many of us became skeptical of buying cars that were too complicated to repair on our own. Over time we came to understand that our cars are much safer, more fuel efficient, and more reliable than they ever were before. That level of confidence, however, is predicated on a systems approach that includes independent oversight of automotive safety and heavy sanctions against willful misconduct in the industry. I think we need to take a similarly pragmatic approach in this advancing age of genomics and computers.

The “open data” or “open science” movement, spearheaded by such organizations as the National Institutes of Health (NIH), makes one of its key objectives the free and open access to the raw data used in studies. Open
access gives the larger community the opportunity to critically evaluate a piece of published research from the ground up. Under this paradigm, there is no hiding behind fancy mathematical tricks because clinicians, biostatisticians, and data scientists alike all have the chance to find the flaws (and new opportunities) in any particular piece of published research. You may very well soon find yourself in an era where you may not personally be able to verify the validity of a piece of published research and yet we, all of us together, can.

**Question 5:**

In your opinion, who should own the rights to the repositories of data? Health systems, electronic medical record software companies, the public?

**Response from Dr. Macedonia:**

*It really depends on the research. There is no single best answer for this. In some cases there are relevant state or federal laws that govern information ownership. In other cases there must be “data sharing agreements” in place. On a purely philosophical perspective, I am not at all convinced that anyone should “own” data because by its very nature it is something that is emitted and ephemeral. Analytics, on the other hand, represent something created through intellectual effort the same way a song or a work of fiction is created and thus exists as intellectual property. You can count me as someone who is in favor of the open science movement.*

**Question 6:**

You mention that cost is one of the limitations of randomized controlled trials. Are big data analytics cheaper to perform? There seems to be a significant cost to obtaining EMRs, salary time to record all the data, software to analyze the data, and a computer powerful enough to perform the computations. Is there any published information to suggest that one research method is more cost efficient than the other?

**Response from Dr. Macedonia:**

*A significant driver of cost of any clinical trial is patient recruitment. Designing trials that reduce the number of study subjects (using tools like Bayesian adaptive study design) or performing big data analytic studies repurposing existing data are on their face more cost effective, leading most pharmaceutical companies to push for their adoption on this very point.*

*The market analytics firm McKinsey & Company, respected consultants to over three quarters of Fortune 500 companies, stated in their 2013 report that the U.S. could realize a 450 million dollar annual cost savings by adopting big data analytics to health care. Admittedly, the very real costs of the analytics infrastructure you mention are a barrier to some institutions and researchers, but as the infrastructure supporting EMRs expands, those marginal costs dissipate. As mentioned in the article, we have certainly seen this with sequencing costs (mostly through innovations in bio-informatics) and the explosion of cancer-related genomics research.*

**Question 7:**

You comment that journals may be reluctant to publish big data papers, thereby creating bias against young innovators wishing to pursue these methods. Why would journals be reluctant to publish manuscripts based on big data analytics?

**Response from Dr. Macedonia:**

*Before I was a physician I was a chemist. There is a concept in physical chemistry called “initiation.” Take a wooden match and hold it in your hand. You can stare at the match for a hundred days and it is not going to light on fire. Strike the match on a high-friction surface and put a small amount of energy into the match and it releases its energy. Striking the match initiates a reaction taking a stable system, briefly making it unstable, and initiating the reaction that produces fire and light. It takes energy, it involves some instability, and*
sometimes it doesn’t work. To get journals to accept big data papers is going to take energy and certainly involves a temporary period of instability as we bring young investigators into the fold, change our review processes, and accept the fact that there will be some failures. For this reason, I am so thankful to the Editors for their efforts to get this discussion started and introducing the readership to something both new, and for a period of time, unsettling.