

MAHA Grand Study

How to Collect Data that will show correlations between various causes to various diseases and issues

Presented as a thought experiment
but if HHS or a college wants to implement it, go for it.

Tom Currie, September 22, 2025

From the “Musings from the Chair” series of essays

Abstract

- There are lots of claims about the health issues in the US
- There are a lot of credible studies and social media claims for the average person to sort through
- There is no “one stop” data set to indicate credibility of various claims
- Gathering a massive data set from all over the nation would allow analysts to find correlations between the various claims and citizens actual perceptions of their diet and health.
- A massive data set can help separate variables or find issues with more than one root cause
- This proposal shows one way to gather and analyze that massive data set

Introduction

- HHS under RFK Jr is working hard to Make America Healthy Again (MAHA)
- They are looking at many things <get the list from HHS>
 - Autism, chronic disease, allergies, ADHD,
- These issues are very complex and may have hard to see connections or dual root causes. <dual causes could be the most important item the data could reveal>
- Influences: location, environmental, health, diet, excersize, mood, social, on line habits, other?
- Studies and data are being done by highly competent scientists.
- There is a great resource to be tapped to collect and analyze data (COLLEGE STUDENTS)
- This proposal sets up a system that can get a LOT of DATA and analysis quickly
- This proposal is intended to be improved by others as it unfolds

Problems

- The 5 A's
 - Anxiety
 - Allergy
 - ADHD
 - Asthma
 - Autism
- Other
 - Celiac (gluten sensitivity)
 - Diabetes
 - Obesity
 - Chronic illness
 - Etc

Possible Causes of Chronic Disease

- Causes of Chronic Childhood disease, per MAHA commission
 - ultra processed food
 - environmental toxins
 - chronic stress
 - inactivity
 - over medicalization of children
- But these are somewhat vague

What if an issue has two causes?

How would you sort that out?

- Hypothetical thought
 - What if Celiac (Gluten intolerance) is actually caused by
 - ultra processed food & environmental toxins
- Another example (hypothetical)
 - What if a chronic children's disease was caused by both
 - ultra processed food & over medicalization of children

How to Indicate the Root Causes?

- Specific, detailed scientific measurements can be done by the medical community
- But before determining the chemistry to study, you need to indicate what to look at.
- PROPOSAL:
 - Take a lot of data by interviewing tens of thousands of people
 - Use a standard list of questions
 - Create a database that is publicly available
 - Analyse the data for correlations
 - e.g. diabetic children tend to share these diet, activity and medical traits

The Great Resource: College Senior Projects

- There are almost six thousand colleges in the USA
- Typically each senior needs to do a “Capstone Project” or equivalent
- College students have a wide variety of talent
 - Social workers, software engineers, web site designers, AI engineers, medical students, statistics majors, chemists, engineers, aggies, managers, etc
- Leverage the talent stack to create a “Grand Study” of the nation
- Leverage the fact that students have a 3-6 month window to execute and finish their project
- Leverage multiple colleges to get various aspects done concurrently

MAHA Grand Study Outline

- Phase 0: no need to identify target problems, add a variety of problems for the database to cover, i.e. all we can think of with room for folks to find new correlations, not just targeted ones.
- Phase 1: Foundation
 - Setup all the stuff needed to collect and save the data
- Phase 2: Data Collection
 - Send people out to collect needed data and enter it into the database
- Phase 3: Data Analysis
 - Do analysis on team, college, regional and national data. Record observed correlations
- Phase 4: Vet Results of Analysis
 - Each conclusion should stand up to scrutiny of various subject matter experts and additional data.
- Phase 5: Determine additional scientific research is indicated

Phase 1 Foundation

- Determine what to look into from MAHA and national discussions
- Sketch out the system: management, development, data collection and analysis, what teams are needed?
- Create a questionnaire
- Create a searchable database to accept data from many sources
- Create a website or app to facilitate data input
- Create a plan to mobilize thousands of data collectors
- Create a plan to do initial analysis of the data
- Create a plan to do data analysis of local, regional, and national results
- Create a plan to interpret and vet the analysis
- Create a plan to act on the analysis
- List the fields of study needed for all this

Phase 2 Data Collection

- Form teams at multiple colleges in every state
- First level collection: ask college students
 - Collect data on thousands of college students
 - Use this as initial data
 - Debugs the database and questionnaire
 - Provides good early data and results
- Second level collection: the public
 - Sample all walks of life outside of college
- Get data from key demographics, organic families, junk food families, ethnic food eaters,
- Enter data into the “Raw Data” database

Phase 3 Data Analysis

- Each college team does their own initial analysis
 - Find problems and resolve them
 - Get early results, draw conclusions on the limited data
 - Each student gets a more complete and useful view and education
- Other college teams do analysis on larger data sets
 - All data from one college
 - Regional data
 - Statewide data
 - National data
- Enter results into the “Analysis” database

Phase 4 Vet Results of the Analysis

- Various disciplines review the results of the conclusions
 - Check the ones that match expectations
 - Check the ones that surprised us
- Are the results reproducible across regions?
- Are results properly interpreted?
- What results need to be re-examined?
- Enter the results of the vetting into the “Vetting” database

Phase 5 Determine what Research is Indicated

- Review the results of the analysis and determine what research experiments are indicated
- Specific chemical or nutrition studies may be indicated
- Specific combination experiments may be indicated
- Initiate a whole new list of College Level Senior Projects

Questionnaire Suggestions

- It needs to be very detailed and easy to fill out (maybe on line via website)
- It needs to have a standardized format for easy entry in a database
- It needs to have a check to ensure data is not faked
 - each questionnaire tagged to a project team ID with verifiable students
- Ask about a very wide range of influences on life, per person in household
 - Age, hometown, current city, durations lived there, vaccine history, diet descriptions, medical data (illnesses of note, etc), hobbies, sports,
 - Add some less obvious influences, like intellectual pursuits, reading, music, art, assembly projects, sports (team, individual, shooting), is the person generally happy/neutral/sad/mad? Number in the friend circle (same age friends, multiage friends)
 - Example: shooting sports have lead, smoke, ventilation (indoor/outdoor), mind focus (safely and aim), body control (aim, cowboy gun events, skeet), exercise (hunting)
 - Include questions that are unexpected/unrelated. Allow others to determine if that data point is useful.

Database Suggestions

- Databases need to be online
- Databases need to be uniform (one database for raw data)
- Need four databased: Raw Data, Analysis, Vetting, Indicated Research
 - Analysis database may need two sections: participant analysis, public analysis
- Every team across the country needs to enter data in a simple fashion
- Databases need to be searched and analyzed by all teams and entire public
- Databases need to prevent false data entry
- Databases need to prevent “denial of service” and other attacks
- Databases need to be searched by causal users (the public, students, new team members)
- Databases need to hide identifying information. Don’t let stalkers find real people.
- Databases should prevent data changes, but entries could be marked “invalid” for exclusion or to be ignored if new results invalidate old analysis results

Data Collection Suggestions

- Find the balance between anonymous data and followup interviews
 - May need to disallow followups, but if data is messed up or more questions are indicated...
- Balance a detailed questionnaire and easy of filling it out
- Have the students fill in the data in an app or webpage
- Find a way to get simple answers and background
 - limited keywords vs details added in prose
- Need Family data and family member data
- Collect data for hundreds of college students immediately (over 3 weeks)
- Collect data from the public quickly (over 3 months)
- Collect data from Hundreds of thousands of people, cities, suburbs, colleges, military, etc
- Does each individual questionnaire go into the database immediately or get reviewed? Maybe each team has a reviewer working concurrent with the data collection.
 - Maybe teams of 3, two collecting data, one reviewer?
- Enter data into the database promptly, or automatically
- Collect data in all types of locations
 - Can't just collect data in safe colleges and suburbs
 - How do you collect data in the more dangerous neighborhoods? Maybe bring participants in for a pizza/data party?

More Data Collection Suggestions

- Each questionnaire has an optional field for contact information for follow up
 - OK to leave blank, used for those who feel their situation is unique or acute
 - Data analysis can filter out all who leave it blank or all who fill it in
- Data Analysis must be able to filter by data field
 - Filter by region, type of neighborhood, contact info filled in or blank, questionnaire filled in on line or via interview, etc
 - Allows analyst to look at portions of data, block suspect data, etc
- System must have a way to prevent bots political activists or jokers filling in sour data for sinister or political reasons.
 - Examples of invalid questionnaires
 - If filled in by a bot/robot it is automatically invalid because there is no human described by the data, how do we exclude them?
 - If filled in by a political activist and falsified to skew data, it is invalid due to fraud
 - If filled in by a child without supervision, the well intended data set is likely invalid. Can have a field for age.
-

Questionnaire Rules

Use the honor system, do all we can to exclude invalid data

- Questionnaires filled in via only a few methods
 - Official interviewers
 - On line contributors
- All respondents must pledge to fill it out honestly, without bias to the results or their perception of a scientific reality
 - Let the data reveal the results
- Respondents should fill it out thoughtfully and give their best answer in the moment.
- Optionally, for a more accurate data set, read all the questions, prepare your answers, then fill out the questionnaire
- Skip any answers that are too personal, the answer is unknown or the question is confusing.
- The data needs a multitude of honest respondents to overwhelm the mistakes and errors.
- Have fun, you are contributing to a better understanding of our complex world and how our food, lifestyle and medical actions interact.

Questionnaire and Data Set Realities

- Most of the data will be valid
- A fraction of unknown size will be invalid
 - Bots, activists, children, duplicates, jokers, poorly informed respondents
 - Some of this can be filtered out at submission
 - Some can be filtered out at analysis
 - Some will live in the database. It impacts the margin of error.
- The app or website for submission must add automatic tags to prevent duplicates but perhaps allow corrections by the original respondent
- The data set should be frozen periodically to mitigate denial of service like attacks (bots with fake data).
- There are always humans out there that want to spoil anything fun or useful. The system needs safeguards against them.
- Similarly, well meaning humans make mistakes, need a way to correct mistakes or block incompetent data.
- The data needs a multitude of honest respondents to overwhelm the mistakes and errors.

Validate Prior Claim: Acetaminophen during Pregnancy

If it's science, it's repeatable. The correlation will show up in the data.

- The questionnaire needs one about acetaminophen during pregnancy
- For a mother or father
 - During pregnancy, I (or my children's mother) took acetaminophen this often
 - Never, rarely, monthly, weekly, daily
 - How many kids do you have? Ages?
 - How many of them are on the autism spectrum? (Answer by age)
- For a college student or similar
 - During pregnancy, my mother took Tylenol this often
 - Never, rarely, monthly, weekly, daily
 - I am/am not on the autism spectrum (dare we ask? minimum, moderate, significant)

Conclusion

- Built a large machine that can design an experiment, collect huge amounts of data, do appropriate analysis, indicate new research
- Find correlations in the data to indicate how to Make America Healthy Again
- Find correlations that other scientific methods are blind to
- Allow the public to augment the analysis, there are a lot of smart folks out there.
- If the study and data proves helpful, repeat in other countries (e.g. Japan, Italy, Norway have very different diets), repeat in US in 5 or 10 years
- Next Steps: Muster the students to start the project !