Randomized Clinical Trial Is Biased and Invalid In Studying Chronic Diseases, Compared with Multiple Factors Optimization Trial

Jianqing Wu[1], Ph.D. J.D. and Ping Zha[2], M.D. (Chi. Med.)

Correspondence: tempaddr2@atozpatent.com

1. End the Incurable Era (Independent researcher for cause), P.O. Box 689, Beltsville, MD 20704. www.igoosa.com.

2. Independent Researcher (Not affiliated with any entity), can be reached by the above address.

Keyword: clinical trial, statistical analysis, experimental error, chronic diseases, health optimization, hypothesis test, variance analysis, covariance, stratification, randomization

ABSTRACT

Chronic diseases are still known as incurable diseases, and we suspect that the medical research model is unfit for characterizing chronic diseases. In this study, we examined accuracy and reliability required for characterizing chronic diseases, reviewed implied presumptions in clinical trials and assumptions used in statistical analysis, examined sources of variances normally encountered in clinical trials, and conducted numeric simulations by using hypothetical data for several theoretical and hypothetical models. We found that the sources of variances attributable to personal differences in clinical trials can distort hypothesis test outcomes, that clinical trials introduce too many errors and too much inaccuracies that tend to hide weak and slow effects of treatments, and that the means of treatments used in statistical analysis have little or no relevance to specific patients. We further found that a large number of uncontrolled co-causal or interfering factors normally seen in human subjects can greatly enlarge the means and the variances of the experimental errors, and the use of high rejection criteria (e.g., low p values) further raises the chances of failing to find treatment effects. As a whole, we concluded that the research model using clinical trials is wrong on multiple grounds, under any of our realistic theoretical and hypothetical models, and that misuse of statistical analysis is most probably responsible for failure to identify treatment effects for chronic diseases and to detect harmful effects of toxic substances in the environment. We proposed alternative experimental models involving the use of single-person or mini optimization trials for studying low-risk weak treatments.

INTRODUCTION

Medicine started emerging after the Industrial Revolution in the 18th century. Over the last 150 years, the field of medicine has accomplished some astonishing achievements. Most of them are for treating acute diseases such as bodily injures, infections, poisoning, pains, and trauma, etc. In each of those cases, drugs are not used to restore impaired or lost balance in the body. Despite the success in treating acute diseases, medicine has failed to find cures for chronic diseases. Main evidence for its failure includes:

(1) Nearly half of all adult Americans suffer from at least one chronic disease. This is equivalent to approximately 45% or 133 million of the population; (2) nearly all chronic diseases are officially listed as incurable diseases in medical references. A long list of chronic diseases is still without cure. In addition, many types of cancer are considered as incurable and terminal; (3) in 2009, 7 out of 10 deaths in the U.S. are attributed to chronic diseases. Heart disease, cancer and stroke account for more than half of all deaths each year. We estimated that the total number of premature annual deaths attributed to chronic diseases is about 30 million in the world based on total death data [Tinker, 2014; Fried, 2017; Raghupathi and Raghupathi, 2018].

The failure of finding cures is best reflected in cancer. A systematic review concluded the complete response of rates of chemotherapy for later stage of cancer have remained static and locked at about 7.4% [Ashdown *et al.* 2015]. The response rate of thyroid cancer treatment was 22.1%-27.1%, with complete response rates being 2.5%-2.8% [Albero *et al.* 2016]. A recent study examined the most promising cancer treatment methods, and concluded: “The claimed ‘targeted’ therapies that may or may not extend remission of cancer for a few months should not be accepted any longer as ‘cure’ by oncologists, scientist or patients….” [Maeda and Khatami, 2018]. The prevalence of chronic diseases in the U.S. has become a huge burden on the U.S. In a study done by the Milken Institute, the annual economic impact on the U.S. economy of the most common chronic diseases is calculated to be more than $1 trillion, which could balloon to $5.7 trillion by 2050 [Milken Institute].

We see that medicine advances on two distinctive tracks. It is capable of achieving astonishing achievements in the treatment of acute diseases. However, it fails to find cures for chronic diseases. The clear separate line between the two kinds of diseases seems to indicate that the performance difference is related to the medical research and practicing models. In this article, we explore if the population-based clinical trial has some inherent

limitations that prevent medical researchers from finding cures.

METHODS

Our purpose is to examine the performance of clinical trials and statistical methods in the context of characterizing chronic diseases.

A. Basic Model Assumptions

We suspect that human individuals introduce very large variances to any measured health properties so that clinical trials are unfit for studying chronic diseases. To prove it, we will use following model assumptions:

Treatment: $s_1 \sim N(\mu_1, \sigma_1^2)$ that affects a trial outcome

True error: $\varepsilon \sim N(0, \sigma_E^2)$

Other causal or interfering factors: $s_2, s_3, \ldots, s_k$.

$s_2 \sim N(\mu_1, \sigma_2^2)$

$s_3 \sim N(\mu_2, \sigma_3^2)$

....

$s_k \sim N(\mu_k, \sigma_k^2)$

$s_2, s_3, \ldots, s_k$ include anything that could influence measured health properties relative to trial outcome. They may be substantial cause factors, independent causal factors, indirect causal factors, covariates (independent factors or confounding factors), etc. The only criteria is that their effects are sufficiently close to the intended treatment so that they must be considered in practice.

In a clinical trial, treatment s_1 must be much larger than the total combined effects of ε and all $s_2, s_3, \ldots, s_k$ so that $s_2, s_3, \ldots, s_k$ can be ignored or treated as part of the error ε.

The model assumption in our study is that s_1 is close to ε and also close to one or all of $s_2, s_3, \ldots, s_k$. For example, in a trial to study a caner treatment, a trial outcome may be judged by observing patients' average survival times. A large number of factors shown in Table 3 are known to affect patients' survival times. Those factors may be traced to genotypes, lifestyle, diets, physical activities and exercise, toxic compound levels in the body, virus infections, gut microbiota, other health problems, etc. It is further assumed that they affect patients' survival times randomly. Each of such factors may

appear in some patients but not in other patients.

Our question is whether a randomized controlled trial can accurately detect the effects of s_1 and what could be done to increase the chance of actually detecting the treatment effect which is similar to or weaker than other causal and interference factors. To answer this question, we used a randomized controlled trial model and a mini optimization trial model to evaluate their respective performance. The basic design of the two types of trials are shown in following Table 1 and Table 2.

B. Two Hypothetical Experimental Models

Model A. Randomized Controlled Trial is shown in the below table.

Table 1. Randomized Controlled Clinical Trial With 3 Randomized Interfering Factors

Treatment Arm (TX: s_1)	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
Other Int. Factors	s_2		s_3		s_4	s_2			s_3		s_2
Control Arm (CA)	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	
Other Int. Factors	s_3		s_2	s_4			s_3	s_2		s_2	

The human subjects are allocated to the two arms randomly. The table shows only one potential way of allocation for illustration purpose. The effects of s_1 on health properties is closer or even smaller than any of those interfering factors s_2, s_3, s_4. For example, s_2 may be exercise, s_3 is dietary adjustment, s_4 is stress management, etc. They affect patients survival times like chemotherapy or other treatment (s_1).

Model B. An Optimization Trial is used where all s_1, s_2, s_3, s_4 causal factors and interfering factors are used as one single treatment package for the purpose of raising total treatment effects.

Table 2. Optimization Trial Design With All Four Factors Used As A Treatment (No Randomization)

Treatment Arm (TX: s_1, s_2, s_3, s_4)	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Control Arm (CA)	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10

In this design, all non-treatment causal and interfering factors (s_5, s_6, ... s_k) must be sufficiently small and thus can bundled into the error term. We

call this design as an optimization trial because as many important factors are used as the treatment to deliver maximum treatment effects. Here, all important causal factors and interfering factors (s_1, s_2, s_3, s_4) that would be identified and used are bundled as one treatment package.

We then evaluate how the optimization trial increases the chance to determine true effects of the treatment package and how to increase the chance of finding cures for chronic diseases.

C. Our Analysis

Our focus is on how to resolve true treatment effect when it is influenced by one or more interfering factors. Our initial focus is the accuracy and reliability required to detect the true effect of the treatment.

1. We examined the machine-repairing model to understand why a population-based method similar to clinical trials cannot be used in diagnosing and repairing machines. Attention is directed to accuracy requirement for repairing complex machines and restoring structural and functional balances in machines. We found that an implied requirement for conducting a population-based trial is that all trial subjects must be "nearly identical units."

2. We explored accuracy and reliability required to accurately characterize chronic diseases. One key fact considered is that chronic diseases development speeds. Slow disease development speeds imply small changes in biochemical processes and organ structure in given times. The slow changing rates and small structural changes further imply that a high accuracy and reliability are required to accurately characterize chronic diseases, as compared those for studying acute diseases.

3. We examined personal differences in light of genotypes, phenotypes, and emotional states, or treatment-relevant factors such as race, personal genotype, age, sex, diet, lifestyle, medication use condition, etc. In addition, we further examined massive differences in health properties found in reference ranges of laboratory tests for human beings to determine whether humans can be treated as "nearly identical units". This determination was made in light of high accuracy and reliability required for accurately characterizing chronic diseases. Many aspects of the massive personal differences work like causal and interfering factors on disease outcomes.

4. We determined whether variances in measured health properties attributed to personal differences are too high to satisfy the requirement of "nearly identical units."

5. We collected a number of exemplar causal and interfering factors from medical literature to show how they raise the variances of measured

health properties. We paid attention to variances that arise from race, personal genotype, age, sex, diet, lifestyle, physical activity, exercise, medication history, etc because they have been found to be causal factors, risk factors, or associated factors of chronic diseases including cancer.

6. We examined the logic used in statistical analysis methods such as t-test, z-test f-test, Chi test, frequency test, etc. to determine whether they could remedy the flaw that clinical trials are unable to offer minimum accuracy and reliability that are required to accurately characterize chronic diseases. Assumptions used in statistical analysis were examined in light of our model data.

7. We determined that if the sources of variances from other causal and interfering factors are merged into the experimental error term as an apparent error as in case in Table 1, how the raised variances affect hypothesis test outcomes, results in biased results by failing to detect treatment effects on chronic diseases. We then showed a pattern of bias by conducting hypothetical tests using hypothetical data for our model data comprising a weak treatment and at least one interfering factor with similar effects on the measured health properties.

8. We examined whether health properties from different patients can be added up and averaged as in statistical analysis by using a multiple causes and treatments model.

9. We finally made comparative analysis for two models: a randomized controlled trial and an optimization trial. We showed why randomized control trial is invalid for studying chronic diseases, and showed that optimization trials could offer far better chances for finding treatment effects for chronic diseases.

RESULTS

The clinical trial evolving history reveals that most of early clinical trials were used to investigate malnutrition, infections, and wounds (except rheumatism). No effort has been made to understand inherent limitations of clinical trials in the history. We also note that the functional approach used in machine is inherently incompatible with the population approach (C, Sup.). The population approach cannot be used in diagnosing and repairing machines made of different blueprints. A population approach may be used to study properties of only “nearly identical units.” Differences, if any, must not cause any functional imbalance, structural misfits, fuel imbalance, flow imbalance, heat imbalance, etc. The population approach has not been used

to address mechanical problems.

Whether a health problem can be studied by clinic trials depends on the purpose of the study. A threshold requirement is that the effect of treatment's on health property is sufficiently larger than the experimental error. This requirement can be satisfied in cases studying strong treatment effects such as pain-killers, surgery, antibiotic drugs, sedative drugs, etc. In those cases, differences among persons will not significantly alter results.

A. High Accuracy and Reliability Required for Studying Chronic Diseases

"Chronic diseases are defined broadly as conditions that last 1 year or more and require ongoing medical attention or limit activities of daily living or both." [Raghupathi and Raghupathi, 2018]. We show the level of balance required in a human body is much higher than the degree of matches between parts in a machine. Health problems can arise from small biochemical imbalances, which result in small changes in structure, shape, and capacity of body components (A, Sup.). As shown in those examples, the deviations in biochemical and cellular processes for causing chronic diseases are "infinitesimally small." Net departures from ideal numbers are often in a tenth percent to a few percent of the ideal personal number. Most net conversion rates must be of right values, and small departures from ideal numbers in either way can be the cause of chronic diseases.

B. Clinical Trials Do Not Support Accurate Evaluation of Health Properties for Chronic Diseases Due to Massive Personal Differences

The population approach is extended to all areas of medicine, but one problem that has never been studied is personal variations. Two big sources are different genotypes and phenotypes [Ogino et al. 2012; Ogino et al. 2013]. The chance of match between two unrelated persons is like that of a DNA match (1 in 113 billion based on 9 loci; 1 in 400 trillion in 13 loci). In addition, personal differences further arise from different emotional states. The personal differences that are important to health may be expressed, in the alternative, as diet, lifestyle, emotional state, culture, environment, sex, medication history, etc.

Personal differences are reflected in reference ranges of laboratory tests for human beings, which are established by empirical methods. The

reference ranges, which reflect measured variances in any health properties in a population, depend on personal differences in genotypes, phenotypes, daily fluctuations, and measurement error. Reference ranges of more than five hundred health properties are published [AccessMedicine]. The measured value of each health property for any person will fall a distinctive point of the range shown in Table S1 (D, Sup.). If each reference range is divided into N levels which could be resolved by detection resolution, the total number of variants of all health ranges would be the product of all possible numbers of all reference ranges. It could be infinitely large. Each of the health properties of a person may fall a distinctive position of the corres pondent population's refernce range. No person would have all of his health properties match the population's means.

All departures of a person's measurements of health properties from the population's means are necessary to correct genetic weakness or to maintain the phenotype, and thus are presumed to be important in maintaining health and prevention of diseases. Differences between two indi vidual persons can be inferred from differences in body size/shape, organ siz e/shape, structural strengths, skin colors, physical capacities, emotional con ditions, etc. Differences are also reflected in diagnostic data, image data, he alth conditions, disease histories, etc.

Personal differences must be considered in treating chronic diseases. F irst, when persons are sufficiently different, they cannnot be treated as same or similar units in a clinical trial because their differences can interfere with the measured health properies in the trial. Second, the values of health properties cannot be used as parameters for predicting chronic diseases. Such health properties cannot be correlated to conversion rates of metabolites and net size on tissue structures. Health properties may fluctuate in daily, weakly, monthly or yearly basis within the lowest and highest ranges. Chronic diseases may arise when health properties in a person depart from optimal values for sufficient duration. Cures to such diseases would require correcting such small departures. Finally, personal differences, which is reflected in health properties shown in Table S1, affect both disease process and healing process. Personal numbers such as vitamins levels, heavy metals, HDL, LDL, cholesterol, platelet count, red blood cell, white blood cell count, fatty acids, glucose levels, triglycerides, etc. can be altered by changing a large number of lifestyle factors.

A review of the history of clinical trial development history, we show that none of old studies we could find discussed personal differences, interfering factors, and their effects on a weak treatment effect (B, Sup.). In a traditional clinical trial, the treatment effect is much stronger than the experimental error so that interfering factors will not alter trial outcome (Figure 1). Absolute reference in our figures is an imagined health property

that could be measured when the treatment is not applied. Since a chronic disease is caused by changed balances, the absolute reference is the disease state when those deviated balances such as excessive omega 6/3 fatty acid ratio, excessive heavy metal levels, physical inactivity, abnormal gut microbiata, lack of dietary fibers, abnormal emotional state, etc. are not corrected. An absolute reference exists in a patient, but could not be applied to a population. It may be used to a small number of "sufficiently similar patients" only if research focus is limited to a small number of interfering factors.

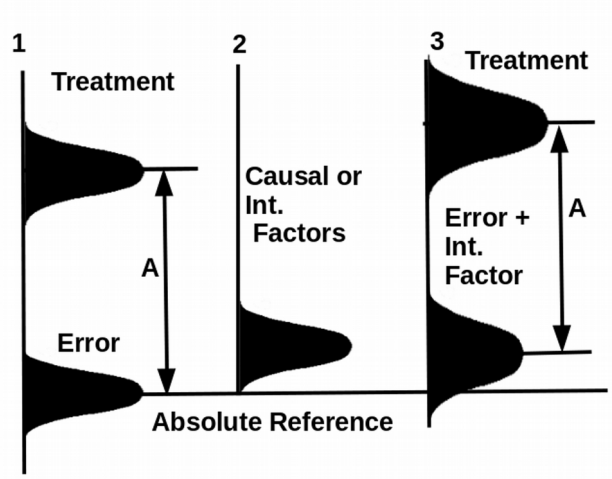

Figure 1 Treatment Effect Is Much Larger Than The Sum of an Interfering Effect and Error.

In studying a strong treatment effect (Figure 1), an assumption can be made that all persons can be treated as identical units because the strong treatment effect cannot be distorted by interfering effects in meaningful amounts. Any differences caused by personal differences are so small that they can be properly neglected. In this situation, randomization is sufficiently good. The justification of use of clinical trials is good approximation. After the error and interfering factors are summed up, resulting in a new distribution under the line 3 (E, Sup.), the treatment effect is still much stronger than combined effects of the error and the interfering factor. Even if many interfering factors exist, their effects could still be neglected.

In studying a chronic disease (Figure 2), the treatment effect is weak relative to two interfering factors shown under the line 2. When the two interfering factors and the error are summed up, they generate an apparent error distribution under the line 3. The mean of this apparent error are the sums of the means of the error and means of the two interfering factors. Without considering the interfering factors, the trial is to find the differences between the treatment and the error under line 1. If the interfering factors

are considered, the trial actually determines the treatment effect under the line 4 relative to the apparent error under line 3. The trial may be unable to find the treatment effect if the data comes out with the treatment's effect at a lower tail region and the error at the upper rail region.

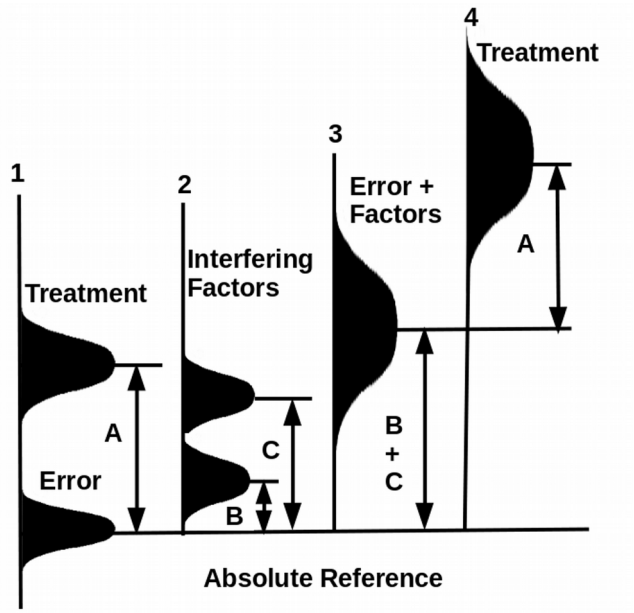

Figure 2 Treatment's Effect Is Close to the Sum of Error and Two Interfering Effects.

In a worst situation (Figure 3), the effect of one or more interfering factors is larger than the effect of a treatment. In this case, the error under the line 1 and the interfering factor under the line 2 merges to become a large apparent error with large variances under line 3. The treatment and the apparent error have a large overlap region (if the profile under 3 is moved onto line 4 horizontally). A trial may come out with the treatment effect falling at the lower tail region while the apparent error at the upper tail region, resulting in a finding that treatment is negative relative to the control. This result is clearly against the model assumption that the treatment has a weak effect indicated by letter A.

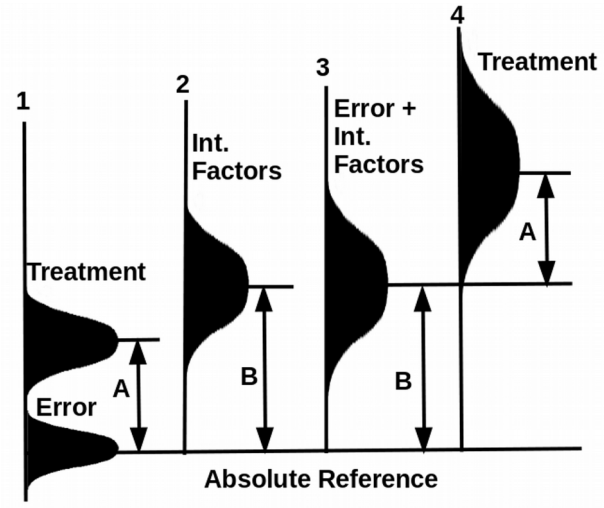

Figure 3 Treatment's Effect Is Smaller Than the Sum of an Interfering Effect and Error.

When the weak treatment overlaps the apparent experimental error as shown in Figures 2, and 3, the trial is meaningless. Nothing can correct this problem that arises from breaching the basic presumption that the treatment effect must be much larger than the experimental error.

Figure 4 shows how an optimization trial by including the interfering factor (which appears in Figure 3) as part of the treatment will dramatically improve the chance to determine the treatment effect. Optimization with both the original treatment and the interfering factor will reduce the variances of the apparent error and increase the difference (designated by A+B) between the mean of the whole treatment package and the control.

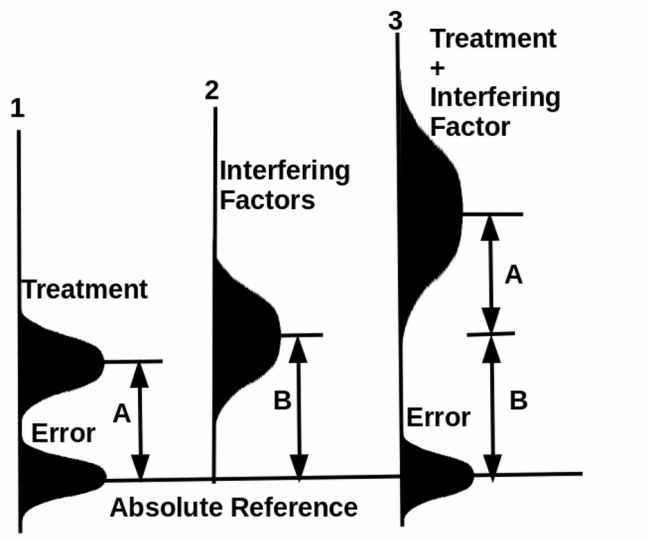

Figure 4 Treatment's Effect Is Increased by an Interfering Effect When An Interfering Factor is Used as Part of the Treatment in Optimization

In studying a chronic disease, all persons must be regarded as different. Cancer provides best example in this regard. Each tumor is unique due to the genetic and epigenetic basis and exogenous exposures such as dietary and lifestyle factors [Ogino et al. 2013]. If a treatment protocol developed from the population data can be used to cure the disease of a particular person, one would have to wrongly argue that the health properties of the person are unimportant to diseases, and phenotypes can be freely changed. Health properties are not quantities that can be summed up and averaged, and a treatment protocol based on population data cannot be applied to any specific person as cures for chronic diseases. This might be the reason why medicine could not find cures by using clinical trials.

If a statistical analysis of clinical trial data yield a "significant difference" over a large number of interfering factors, such a treatment must be very strong. It could be unlikely for such a strong treatment to correct weak causal causes for chronic diseases. This might be a reason that treatment protocols from clinical trials are able to control symptoms quickly, but are unable to restore sophisticated balances in human bodies.

C. Multiple Co-Casual and Interfering Factors Are Prevalent in Human Beings, But Often Could Not Be Controlled in Clinical Trials

Massive differences among individual persons are anticipated to affect the accuracy and reliability required for studying and characterizing chronic diseases. In a large clinical trial, a measured health property such as survival time or hazard ratio depends on disease progression, the effect of the treatment, all uncontrolled interfering factors, and their interactions. Naturally, all those factors are added into the error term. The final conclusion of the trial depends on the treatment effect relative to the bloated error term. If many factors are not controlled, the presumption that the treatment effect is much larger than the experimental error fails and result is incorrect. We will show a list of uncontrolled factors that can be seen in clinical trials.

Table 3. Factors That Are Known to Influence Chronic Diseases; Most of

Them Normally Are Not Controlled in Clinical Trials

Case No.	Classes	Impacts and Mechanisms	Implied Effects and Significant Degrees	References
A1	Genetic (mutations).	Cancer initiation, development and metastasis.	As well accepted somatic evolution theory.	[Nowell, 1976; Loeb, 1991]
A2	Genetic (agiogenesis)	19 Endogenous angiogenic polypeptides (VEGF, APN, etc.).	Tumor mass is limited to 1–2 mm without neovascularization.	[Sagar *et al.* 2006]
A3	Genetic (apoptosis)	Inflammation, apoptosis and autophagy.	54 genes related to those properties in breast cancer.	[Schuetz *et al.* 2019]
A4	Genetic (race).	Among Asian people: Stroke is more prominent than CHD.	Different characteristics.	[Ueshima *et al.* 2008]
B1	Age and aging on cancer incidence.	Cancer incidence rate is related to sixth power of age.	Huge impacts on prevalence and cancer death rates.	[Armitage & Doll, 1954]
B2	Age on inflammation..	Aging and hormonal changes.	Age has great impacts on inflammation.	[Prasad et al. 2012]
C	Sex.	Males had higher age-adjusted death rates for 12 of the top 15 causes of death.	Sex's effects depend on diseases.	[Kalben et al. 2000]
D1	Chronic stress.	Cancer initialization, growth and metastasis.	Affect immunity; neuroendocrine/β-adrenergic signaling.	[Segerstrom and Miller, 2004; Sloan *et al.* 2010]
D2	Chronic stress.	Increased levels of platelet–leukocyte		[Brydon *et al.* 2006; Sundquist *et al.* 2005;

		aggregates.		Nemeroff *et al.* 1998]
E1	Diet (vitamins, fibers, minerals, etc.).	Nutrition affects tissue ecosystem and cancer proliferation.	A massive number of studies show diet's impacts on various aspects of health and cancer.	[Ogino et al. 2012, 2013] (too many to include).
E2	Diet (natural compounds) .	Targeting apoptosis pathways in cancer.	A massive number of natural compounds.	[Millimouno *et al.* 2014]
E3	Diet and gut microbiota	An impaired microbiota dysbiosis linked with cancer.	Probiotics, potential corrective diet, Fecal Microbiota Transplantation	[Vivarelli et al. 2019]
F1	Systemic inflammatio n	Affect the tissue ecosystem and inflammation.	Through age, body mass index, dietary saturated fat, and EPA and DHA omega-3 fatty.	[Navarro *et al.* 2016]
F2	Diabetes and Cancer.	Potential diabetes and cancer association.		[Giovannucci *et al.* 2010]
F3	Viruses	DNA viruses and RNA viruses	Cause of about 15% of all cancers in the world.	[Liao 2006]
G	Prior chemo.	Cancer repopulation.	Mechanism is unknown.	[Salani *et al.* 2011]
H	Surgery.	Cancer repopulation.	Systemic inflammation; promote M2 Tumor Associated Macrophages.	[Krall et al. 2018; Colleoni *et al.* 2016; Cheng *et al.* 2012; Demicheli *et al.* 2007]
I	Radiotherap y.	Tumor cells repopulation.	Mechanism is unknown.	[Salani *et al.* 2011]
J1	Exercises.	Cancer initiation, growth and	28%–44% reduced risk of	[Cormie *et al.* 2017]

		metastasis.	cancer-specific mortality.	
J2	Exercises.	Affect 35 Chronic diseases.	Good rehabitative therapy.	[Booth *et al.* 2012]
K	Temperature, vibration, etc.	Enzyme activity; cell division apparatus.	(Influence cancer by time-averaged effects.)	[Levine & Robins, 1970] [Yeung *et al.* 2003]
L	Lifestyles	Association of risk factors to myocardial infarction.	Changing lifestyle could prevent at least 90% of myocardial infarction.	[Yusuf *et al.* 2004]

The above table shows only a few exemplar factors that normally influence chronic diseases including cancer. The exact working mechanisms are unimportant to our analysis. Those factors affect a treatment result for a chronic disease if the treatment is evaluated by measuring a health property such as survival time, hazard ratio, chemical analysis data, structure's size, biochemical process speeds, etc. They affect measured health properties by causal effects or by influencing one or more causal factors. Some factors may work like confounding factors.

Variances of each factor arise from an error in measuring the factor and the mechanisms at which the factor affects the measured health property. For example, it is impossible to accurately measure intensity, amount, and duration of exercise. Even if exercise were used in a precise accuracy, actually delivered effects on the health property would depend on personal conditions.

Surgery is considered as a factor of influencing cancer cell re-population by different mechanisms. Exercise is found to be an important adjunct therapy in the management of cancer based on a large number of studies (Cormie et al, 2017). Physical inactivity is one important cause of most of 35 chronic diseases [Booth *et al.* 2012]. Chronic stress can dramatically speed up cancer metastasis [Segerstrom and Miller, 2004; Sloan *et al.* 2010]. A prior surgery can dramatically alter the body's ability to resist cancer return growth speed [Krall *et al.* 2018; Colleoni *et al.* 2016; Cheng *et al.* 2012; Demicheli et al. 2007]. Age affects cancer incidence rate by a sixth power [Armitage & Doll, 1954]. Age, body mass index, dietary saturated fat, and EPA and DHA omega-3 fatty can affect the body's inflammation potential [Navarro *et al.* 2016]. Many uncontrolled factors may be magnitudes stronger than treatment's effects when their effects are

looked in long terms.

In clinical trials, most of those factors are not controlled or cannot be accurately controlled. For example, surgery cannot be well controlled. If patients in a typical trial have been operated previously, the amount of tissue loss and surgical locations are dictated by medical needs. Ages may be classified by age groups but their effects cannot be well controlled due to personal differences. Most lifestyle factors cannot be measured accurately and thus are anticipated to have different effects. Since people have different lifestyles, their prior lifestyles may have residual effects on health properties after their lifestyles are changed per required treatment.

If a clinical trial is designed to study a weak factor, tens to hundreds of other uncontrolled factors with similar levels of effects are "bundled" into the error term. All of those factors affect human subjects in both the treatment and the control; and due to randomization, they do not cause meaningful difference between the treatment's mean and the control's mean. Each interfering factor raises both the mean and variances of the apparent experimental error term (See Figures 2, and 3). We will show that statistical analysis not only fails to correct such a problem, but makes the problem worse by failing to recognize weak treatment effects.

D. Randomization and Statistical Analysis Cannot Correct Bias Caused By Personal Differences, but Increases the Chances of Accepting Null Hypothesis

Randomized control trial does not automatically deliver a precise estimate of the average treatment effect, and it yields an unbiased estimate, that applies only to the sample selected for the trial [Deaton and Cartwright, 2018]. They discussed many problems, but did not discuss the inherent bias when a treatment effect is weak while multiple interfering factors exist. Accordingly, no attempt has been made to understand the merit of using multiple factors optimization method.

One common type of statistical analysis is to compare the mean of a treatment with a control by conducting a hypothesis test. Our simulation shows that the statistical outcome depends on data dispersion. In cancer cases, if the survival times become more widely dispersed, the point for rejecting the null hypothesis will shift toward to a high value. This means that a weak treatment effect will be rejected as random errors at high chances (see all examples in F-I, Sup.). This can be seen from Figures 2,3 as well.

In conducting a hypothesis test by using t distribution, a health property is observed before a treatment and after a treatment. The paired difference is used in conducting a hypothesis test. The rejection point depends on how patients respond to the treatment similarly. If they respond to the treatment in the exactly same way, even a small treatment's effect can be recognized. However, large differences in patients' responses will cause the rejection point to move toward to a large value for the same p value, and thus fails to recognize the effect of the treatment (F, Sup.). In conducting two populations' mean test, large differences within each treatment group will cause the rejection point to shift toward to a large value (G, Sup.).

In conducting a variance analysis, uncontrolled interfering factors affect the health property to be measured. The test outcome depends on ratio of the variances of the treatment to the variances of the random error. If interfering factors are not controlled, they will go into the error term and thus reduce the ratio of treatment variances to error variances. The uncontrolled factors cause F statistic to shift to lower value so that the F test will be more likely to accept the null hypothesis (H, Sup.).

Interfering effects of uncontrolled factors cannot be corrected by any other statistical analysis method including $\chi 2$ goodness-of-fit test, common frequency fest (J-K, Sup.). Some statistical methods take into account only sampling drawing error, and others may address specific problems, but none have the power to correct this fundamental flaw that must be addressed by raising measurement accuracy. The problem cannot be cured by any methods such as randomization and stratification (L-M, Sup.). Simpson's Paradox is also powerful proof that different persons cannot be treated as same in a clinical trial (N, Sup.).

Prior studies on the benefits of randomization in clinical trials are focused on how randomization reduces systematic bias [Kalish and Begg, 1985; Fleiss et al, 2003] and prevents selection bias [Schul and Grimes, 2002]. When human subjects are randomized, all [interfering] factors that affect experimental outcomes can be similarly allocated to the treatment group and the control group. This similarity allows for statistical inferences on the treatment effects [Altman, 1991]. While those points are correct in the context of studying a strong treatment as shown in Figure 1, they did not consider the effects of uncontrolled interfering factors when their effects similar to the treatment's effect. They did not consider how combining multiple factors as a single treatment can dramatically raise the capability to detect treatment effects.

The root cause of Simpson's Paradox is large variances at personal levels. In characterizing a chronic disease, each person must be treated as a unique system. A distinctive regression curve is presumed to exist for each person. When data from different people are pooled in conducting a

regression analysis, it is an attempt to find a regression curve among different systems. Such a regression curve cannot be right except by accident. It may be applicable to a population, but the population does not have diseases. Thus, any treatment developed using population data cannot cure diseases for any specific person. The pattern of Simpson's Paradox implies that such a regression data is improperly combined.

The problem discussed is rooted in the fact that massive personal differences in clinical trials affect a measured health property. No statistical analysis, and nor any other methods under the Sun can correct this problem, which is like a bad laboratory report which is based on data generated by using an erratic household scale. The causal and interfering factors include health factors that patients can correct and factors that patients cannot change. Some of interfering factors are called as covariates; and some examples include sex, age, trial site, disease characteristics, disease prognosis, etc. [CHMP, 2015]. A presumed fix is to achieve balance among treatment and control arms with hope that the conclusions of a clinical study are not sensitive to covariates. However, none of the proposed methods in the Guideline can actually correct the bias of clinical trials because those measures cannot reduce the variances of the error term. In another study, attempts have been made to evaluate different methods for correcting baseline imbalances [Egbewale et al, 2014]. They focused only pre- and post-treatment scores and how different analytical methods affect bias, but did not address the problem of weak effects. The problem cannot be addressed by co-variance analysis.

We also show that health properties are not the types of things that can be summed up and averaged (O, Sup.). Good personal health is maintained by maintaining sophisticated balances. Beneficial effects and adverse or negating effects happen in different patients, and they cannot be averaged in reality. This unique problem arises in the context of characterizing chronic diseases. It is safely assumed that chronic diseases are caused by imbalances, which can be caused by disturbance in two opposite directions. Each biochemical pathway must be maintained at a proper speed, and changing the pathway's speed in either way can disturb this balance. A same amount of qualitative change from a right pathway speed in one person cannot be used to compensate for the same amount of change in an opposite way in another person. However, statistical analysis is based on an assumption that health properties is fungible and thus can be summed up and averaged. This assumption cannot hold in reality. An identical amount of departure from the population's mean has different impacts on different patients. The same amount change may cure, hurt or kill a patient, depending on the specific conditions of the person.

Statistical analysis is based on an oversimplified and unrealistic

assumption that health properties can be treated as a fungible property. Statistical analysis adds negating effects to the sum of the treatment and thus lowers the treatment's mean. This results in wrong result like that a 20% beneficial response rate and 15% negating response rate gives 5% net beneficial response rate, or that the sum of 20% positive effects and 20% negative effects is equal to no effect. In reality, one can avoid negating effects by avoiding applying the treatment to mismatched patients and can actually deliver 20% positive effects. For those obvious reasons, a treatment protocol developed from a clinic trial predictably fails to work on real human beings. Optimization focusing on a single patient is the only way to avoid this fundamental flaw. This problem is less critical when health properties among "sufficiently similar subjects" are summed up or averaged to get rid of fluctuations caused by uncontrollable errors.

Based on a hypothetical model study, where each of k factors can influence the health property by a same amount, using k factors as a treatment is superior to the treatment using a single factor (P, Sup.). If each of the k factors has a same treatment effect and same variances in the health property and is similar to the experimental error, using an optimization trial to optimize the heath property by using k factors will raise treatment effect by k times, and raises the T statistic, Z statistic and F statistic by about k*√k. The sensitivity and ability of a hypothesis test to detect true treatment effect increases with the number of interfering factors. When the total number of factors is increased from 1 to 2, 5, 10, and 100, all statistics will increase by 2.8, 11.2, 32, and 1000 times.

DISCUSSION

A. Multiple Sources of Big Errors in Clinical Trials

In conducting a valid experiment, one fundamental presumption is that accuracy and reliability of detection technologies for detecting a treatment's effect must be sufficiently higher than those for detecting the experimental error. This presumption can hold only in studying strong treatments that can stand out over the apparent experimental error. In medical research, this presumption becomes that detectable treatment's effect must be much larger than the apparent experimental error. In a clinical trial, the failure of this presumption can be rephrased as one that the alternative hypothesis (the effect attributed to a treatment) is too close to the apparent experimental error so that data set tends to come out with its statistic falling within an acceptance region. This results in an outcome of failing to recognize weak treatment effect.

All statistical analysis methods are premised on the model assumptions, and every model assumption including the test hypothesis must be correct [Greenland *et al.* 2016]. The fluctuations caused by beneficial and adverse/negating effects among different patients are not same as drawing error or true random errors in typical statistical models. The effects of uncontrolled factors may be merged into the experimental error only if the total experimental error is still sufficiently smaller than the treatment's effect. Big data dispersion in statistical analysis may not be ignored [Campbell, 1974]. When the experimental error in a clinical trial is close to the treatment's effect, such a trial will generate meaningless results.

Lack of required accuracy and reliability is inherent in clinical trials used to characterize chronic diseases. Chronic diseases, by definition, progress slowly. This means that changes in any measured health property such as hazard ratio, organ function, survival time or other measured chemical data in any given time interval is infinitesimally small. Thus, the accuracy and reliability required to accurately characterize chronic diseases is much higher than those for studying acute diseases.

Compared with mechanical systems such cars, planes, etc, human beings are the most unfit subjects for clinical trials because of a massive number of genetic differences and phenotypes [Ogino et al. 2012; Ogino et al. 2013]. In addition, the personal differences are further increased by different emotional states of human beings. Since the massive personal differences in clinical trials interfere with accurate assessment of any health properties, it is impossible to detect weak and slow treatment effects. By using clinical trials, medical researchers cannot accurately determine what can cure chronic diseases and what harm personal health on long terms.

Statistical analysis has been widely abused in a long history [Campbell, 1974]. Misuse of statistical analysis in medical research is a well known problem which has been discussed in a large number of studies [e.g., Strasak et al, 2007; Gore et al, 1977; Kim et al, 2011; White, 1979; Hall *et al.* 1982]. Problems discussed in those references are in addition of the model flaws we have found above.

B. One-Way Biased Conclusions of Clinical Trials and Their Severe Adverse Impacts On the Global Health Landscape

Our simulation results from all different models consistently show that the effects of clinical trials are one-way biased when the trial is used to evaluate a weak treatment. The averaging operation tends to reduce the treatment mean and this effect is not reflected in any assumption in basic statistical models. Statistical mean, $\mu_{s,}$ must be smaller or much smaller than μ_b the actual beneficial mean when the treatment is used only to cause-matched patients. This effect is described by a degrading factor $g=\mu_s/\mu_b$ which is

attributed to “indiscriminate application” of the treatment. This value is in the range of 0 to 1. Statistical analysis is unfit for studying chronic diseases. If a measured health property is influenced by multiple interfering factors, a study focusing on a single treatment with other factors randomized will increase the chance to reject the treatment as having no effects on the health property. Hundreds trials, with each focusing on one single factor, will result in failure to find any of the factors.

Clinical trials distort hypothesis tests by enlarging the error term and statistical analysis reduces the treatment’s effects by averaging effect. They both work in the direction of rejecting the treatment. If a clinical trial results in rejection of the null hypothesis, the finding will likely stand except that the true treatment effects may be actually larger than determined values. However, if a hypothesis test outcome is acceptance of the null hypothesis, it may be wrong due to the negating effects and interfering effects. Therefore, conclusions in a good number of published studies should be interpreted differently. This one-way bias can be traced to the irreconcilable conflicts among massive personal differences, required high measurement accuracy and reliability, weak and slow effects of treatments for chronic diseases, and the unique roles of imbalances in chronic diseases.

Personal health is influenced by diets, nutrition, exercises, mind regulation, chronic stress, fears, etc. Many of those factors work like double-edged swords: they can benefit some patients, but hurt others if they are misused to destroy some established balances. The effects of nutrition and diets are expected to be highly random and unpredictable due to different personal lifestyles. In such a trial, the apparent error is inflated by uncontrolled interfering factors. Findings from a clinical trial represent only an abstract population, and are inapplicable to real patients as far chronic diseases are concerned. A large number of factors in diet, lifestyles, exercise and emotional states, etc. can affect cancer outcomes, and thus, each study focusing on one single or a few factors will result in rejecting each factor as a potential treatment.

By creating false acceptances, misused statistical analysis keeps rejecting weak and slow treatment effects. This explains why a clinical trial could not positively affirm single lifestyle factor’s curative benefits even though it is found to be a significant risk factor of the disease in other types of long-term studies. Clinical trials are primarily responsible for promoting mainly surgery, synthetic drugs, radiation as “scientifically valid” treatments and rejecting potentially tens of thousands of non-medical weak and slow treatments, which would be one to two orders magnitude more powerful if they are used collectively in optimization trials. Clinical trials are most probably the main culprits that preclude the mankind from finding for cures for chronic diseases. It is reasonable to infer that clinical trials are in part

responsible for creating current national health epidemics in the U.S., China and many other nations in the world.

A serious problem is the cumulative toxic effects of environmental pollutants, contaminants, food additives, pesticide residues, herbicides, industrial chemicals, etc. By focusing on a single toxic agent in each trial, each such study cannot catch a weak and slow toxic agent. However, multiple toxic agents always work together in human bodies. A negative finding could be "caused" by interference of other similar or stronger toxic agents and similar or stronger interfering effects. Most known toxic substances co-exist in human bodies. If a hundred similar factors are studied at the same time, Z statistic, T statistic and F statistic could be 1000 times more than counterpart in the clinical trail focusing on a single factor. When a large number of similarly harmful factors attack the human in the control, each of toxic agents is naturally hidden as "the experimental error." However, several, tens, or even hundreds of toxic agents can slowly damage human bodies. This single toxic agent can be identified only if all those toxic substances are not present in trial subjects. Findings from studying one or few toxic agents a time do not reflect the real damages of multiple toxic agents to the human body.

C. Replacing Clinical Trials by Optimization Trials

To find cure for a chronic disease, a required capability is determining which factors can speed up the disease's progression and which can slow down or reverse its progression. Considering massive differences among human subjects and a large number of interfering factors, clinical trials are unfit for establishing treatment protocols. Optimization trials using multiple factors as a treatment provides much better chances for finding cures for chronic diseases. We will show three huge gains below.

First. the biggest gain from using an optimization trial is to avoid negating effects caused by indiscriminate application of the treatment. For a single factor treatment, an optimization trial can raise beneficial effects by (1/g), where $g=\mu_s/\mu_b$. It can be 1 to any reasonable number (See treatments C to G in Table 7S, Sup.). In clinical trials, the same treatment is indiscriminately used on all patients in the treatment group, many lifestyle factors can disturb various balances in two opposite directions. If those factors are randomly used against all patients in the treatment group or subgroup, their true beneficial effects on some patients can be "nullified" by their negating effects on other patients (per the analysis for the model in Table S7). In an optimization trial, controllable factors are used as part of treatment and are used on only the patients who need them. Sufficiently similar patients are selected in such a trial.

Second, we have shown that a large number of interfering factors

directly interfere with clinical trials. They have different levels of effects and different variances. They can be used as part of treatment package for chronic diseases. Thus, a wise strategy is including multiple factors that would affect disease outcomes as a treatment package. The apparent error distribution in a patient can be estimated by the mean, $\mu_t=\mu_1+\mu_2+....,+\mu_{k,}$ and variances, $\sigma_t^2=\sigma_E^2+\sigma_1^2+\sigma_2^2+,....,\sigma_k^2$ if all interfering factors are not used as part of the treatment. An improvement can be achieved by using the Model B. By bundling all controllable co-causal and interfering factors as a treatment package, the total treatment's effect is raised by k times while the error variances are reduced by about √k according to Medal B analysis.

The total gain in treatment effects existing in an optimization trial over a clinical trial is (1/g)*k while the all test statistic such as T statistic, Z statistic, and F statistic used in hypothesis tests are increased by (1/g)*k*√k, where 1/g is attributed to avoiding negating effects, k is attributed to the additive effect of multiple treatment factors, and √k is attributed to reduction in the error variances. Their collective impacts could be huge. This conclusively shows why medicine could not find "scientific evidence" for any treatment based on a single lifestyle factor.

This conclusion is backed up surprisingly by all of simulation results for three hypothesis tests in every model we used in Supplement. Thus, we assume the gain is an inherent estimate (but not a precise number due to complexity of the human body). Moreover, the actual gain is predicted to be more than (1/g)*k or (1/g)*k*√k. If interfering factors are matched to patients, true gain in the treatment effect is more than k times. We assume that some adverse effects which cannot be directly measured are not reflected in the negating effects and thus could not get into the g value. The true treatment effects could be further raised by avoiding adverse side effects. In contrast, optimization trials are good for using lifestyle factors, natural remedies, and mild or safe environmental factors, they do not implicate serious side effects. Even though the variances of the treatment's mean X could approach zero, the √k term most probably cannot be ignored by approximation in studying chronic diseases.

The inevitable conclusion of clinical trial's invalidity is strongly resonant with ancient medical practices such as herbal formulations and practices under the ancient holistic model. This ancient holistic model stresses the need to work on the whole body by using a large number of natural compounds or multiple treatment methods.

Based on the strength of our evidence as a whole, we reject clinical trails as a misused used wrong experimental method for studying weak and slow health properties and propose optimization trials as replacement. Optimization trial is suitable for studying weak, slow and natural remedies, but may not be used to study the side effects of synthetic drugs.

One solution is using a single human subject in a clinical trial. In this case, a control cannot be found because no two persons are similar in the world. Thus, the person's condition before the treatment is used as a control. This is essentially what was once used in ancient medical systems. The treatment effect is assessed by comparing the health properties before the treatment and after the treatment. One problem is that if the treatment lasts a long time, aging process can interfere with trial results and other previously used treatments may influence the current treatment. Some practical adjustments must be made. Trustworthiness of findings should be established by replicating the same trial for several times. Acceptance of this approach would require examining the rationale of using clinical trials. The notion that a treatment is good for all people in the population is clearly incorrect as far as chronic diseases are concerned.

An alternative solution is controlling all influencing factors in a mini trial so that the variances from interfering factors are minimized. It is difficult to get rid of the massive personal differences which are presumed to interfere with trial outcomes. What could be achieved in practice is using "sufficiently similar subjects" in the mini trial. To investigate a treatment in a trial, all significant co-causal and interfering factors including those listed in Table 3 and other known factors should be controlled. For example, relevant genetics, diet, exercises, toxic agent levels, medication use history, sex, age, race, emotional states, etc. are controlled. When the variances from those factors are controlled, the trial's sensitivity will be dramatically increased. By using sufficiently similar subjects, weak, slow and natural treatment effects can be detected with increased sensitivity. To see whether the treatment works on patients with similar important health properties, a second or third mini trial is conducted. After a series of mini trials have been done, a researcher can see when the treatment works and under what conditions the treatment works.

Personal genetics and emotional states are difficult to control. Personal genetics can be controlled by selecting human subjects. To control those factors, one should focus on their nexuses to measured health properties. If a treatment works on a particular biochemical process, subjects with known genes that control the process should be selected or voided, but other genes with little effects may be neglected. Emotional states should be stressed if they are predicted to play significant roles in influencing measured health properties. When human subjects are nearly identical, variances attributable to personal differences will be dramatically reduced. In personalized medicine, randomization, subject selection bias, statistical analysis has limited utility and should not control experimental designs.

When clinical trials involve a small number of sufficiently similar patients, statistical analysis should not be concerned. When all significant

factors are controlled, measured health properties may be treated as ordinary variables and thus statistical analysis can be avoided or used as a mere causal checks. P-values 0.5±0.15 (or any other suitable numbers) may be used because trustworthiness of trial findings are established by replicating mini trials. For a single person trial, statistical analysis cannot be used in most situations unless study purpose is examining things caused by instruments or sampling technologies, and trustworthiness of findings should be established by repeating the same or similar trial. All details on controlled factors should be documented for replicating the trial.

The single-person or mini optimization trial can be used to study combination of factors. Cancer is clearly responsive to lifestyle changes involving a large number of factors. When tens to hundreds factors are controlled, their combination effects are added up in some ways while co-causal and interfering factors are dropped out from the error term. All co-causal and interfering factors are used to promote healing in the treatment arm. Such an experimental design will dramatically increase the detection sensitivity of the trial and raise the treatment's effect.

When the medical model is flawed, the problems cannot corrected by doing more studies under the same model and its validity cannot be decided by examining studies from the model. Our challenges cannot be evaluated by the same standards under the current standard. To correct such foundation error, one must examine presumptions, medical model development history, past facts that supported the use of the medical model, and newly discovered facts that show flaws of the medical model. The widespread misuse of clinical trials has been driven by the incentive for avoiding selection bias. This incentive is overly stressed in medical literature. This incentive has driven medical research into a wrong track for more than one century, and resulted in a research model barring the use of emotional component as part of cure.

As a result of influences of clinical trials, a notion that diseases are cured by indiscriminate exposure to approved treatments becomes dominate in medicine. The notion is strange. Even though completely response rates of chemotherapy are 7.4% for cancer and virtually no cures are found for most chronic diseases, medicine is still unable to examine its foundation. Despite frequent criticisms by non-medical professionals, medicine is unable or unwilling to examine this flawed research model, and continues using the peer review system to maintain the flawed standard and suppress discoveries that would lead to reformation of the flawed medical foundation. When this fatal flaw is not corrected, medicine continues producing greatly biased, incorrect, or irrelevant research findings. Even after more than a century of failure to find cures and that U.S. medicare is predictably facing bankruptcy, the idea of avoiding selection bias still control research

practices. If an investigator wants to inject bias, nothing can stop him from altering data at trial levels. Besides, a better approach to preventing research fraud and bias is repeating the same trial by one or more times. Other problems such as non-standard definitions, definitions changing in time, inaccurate specification of groups, lack of data, etc. are not fatal and can be addressed over time.

Optimization trials are superior to clinical trials for studying weak effects. By recognizing the validity of single-person trials and mini optimization trials, personal medical miracles can be conveniently studied. Research focus is not about experimental designs, evidence quality, statistical analysis, selection bias, rejection criteria, etc, but delivery of predictable cures which can be tailored to all specific patients including "minority patients." This mission cannot be accomplished by indiscriminate application of treatments in clinical trials.

D. Limitations of This Study

Our findings are not applicable to clinical trials, the findings of which are not used as the basis for treating diseases. If the purposes of research are to explore costs and resource allocations, their validity are not subject to the same analysis. Also, if clinical trials are used to study disease mechanisms as a way to control health costs, they still provide useful information for policy makers.

It has routinely assumed that measured health property in a trial is mainly attributed to a treatment. However, this presumption is always breached in studying chronic diseases. When a weak treatment plus at least one interfering factor affect the measured health property, the validity of trial outcomes depends on the relative size of the treatment to those of the interfering factor. Moreover, concerning chronic diseases, health properties are different from person to person. This implies that true cure must be formulated for each specific patient, and treatment established by population trials cannot restore balance for specific patients except by accident.

We note that the effects of interfering factors are not linearly additive, their effects may vary in degrees, their variances are not similar, their distributions may be not normal, and many factors may interact with each other in complex ways. However, they affect the mean and variances of the experimental error in certain ways. The effects from all interfering factors are added up linearly or non-linearly. When the causal and interfering factors are bundled into the error term, they ruin the trial. If they are bundled into a treatment, test statistics increase as a result of addition of all co-causal and interfering factors, and are further enlarged by an empirical multiplying factor that is attributed to the reduced variances of the apparent

experimental error.

If a clinical trial's design breaches any core assumption, its findings are incorrect for that reason. If the breach is sufficient to change trial outcomes, the trial is invalid without regarding the validity of our findings. Thus, whether or not those assumptions used in our models hold will not affect our conclusions. Our findings underscore the importance of adhering to model presumptions in designing clinical trials and conducting statistical analysis.

CONCLUSIONS

By examining machine repairing model and accuracy and reliability requirements for studying chronic diseases, we found that the one-treatment-for-a-population approach is flawed as far as it is used in studying chronic diseases. Clinical trials are good only if the treatments under study are sufficiently strong or when all human subjects can be treated as "nearly identical units" as in classical probability trials or classical clinical trials. None of the two conditions are met when clinical trials are used to characterize chronic diseases. Randomized clinical trails are unable to deliver required accuracy and reliability due to the massive personal differences attributable to genotypes, phenotypes and emotional states of individual persons.

We further found that clinical trials and statistical analysis are fundamentally flawed on multiple grounds as revealed in numerous hypothetical models such as a multiple causes/treatments model, and multiple interfering factor model, two population means hypothesis test, paired data hypothesis test, F-test in variance analysis, etc. We found that health properties are not the types of fungible things that can be summed up and averaged because all human beings must be treated as different things. Beneficial effects and adverse effects happen in different persons with different meanings, and cannot be averaged in reality. Statistical analysis degrades performance of the treatment by averaging beneficial and negating effects within each treatment or subgroup. This averaging operation dramatically degrades the treatment effects. In conducting statistical analysis, the poor accuracy problem becomes one that the total experimental error is closer or even larger than the treatment's effects under the alternative hypothesis. Both the means and the variances of randomized and uncontrolled co-causal and interfering factors are added to those of the error term as an apparent error. When the apparent "error" is far too large relative to the effects of the treatment, data set tends to come out with test

statistics falling on the region of acceptance of the null hypothesis, thus resulting in false acceptance of the null hypothesis or false rejection of true treatment effects. Those fatal flaws are expected to be present under most circumstances. No statistical method, no any other methods under the Sun, can ever correct this great bias that arises from breaching the core presumption used in the statistical model. Thus, clinical trials are invalid and have been misused in studying chronic diseases.

Our model analysis shows that optimization trials can dramatically increase chances to determine treatment effects than randomized clinical trials. Based on a multiple interfering factor model, where k co-causal or interfering factors can influence a measured health property by a same degree, a treatment package using all k factors is much better than using a single treatment. If each of the k factors has a same treatment effect and same variances, an optimization trial to evaluate the heath property by using all k factors will raise the total treatment effect by (1/g)*k times than a randomized trial (where g is a degrading factor caused by misapplication of a treatment to patients, with its value from 0 to 1), and raises T statistic, Z statistic or F statistic by about (1/g)*k*√k. Assuming that a treatment has no negating effects, when the total number of the factors is increased from 1 (without any interfering factor) to 2, 5, 10, and 100, T statistic, Z statistic and F statistic will increase by approximately 2.8, 11.2, 32, and 1000 times. Moreover, by avoiding negating effects, an optimization trial using k factors as a treatment package can raise treatment effect potentially by one to several orders of magnitude relative to randomized clinical trials. The gain cannot be eliminated by increasing the patient number in the trial. The findings show why studies using clinical trials cannot produce “scientific valid” evidence in support of using single lifestyle factor as cure for a chronic disease.

Clinical trials have be correctly used for centuries but widely misused in studying chronic diseases in the last century. No prior study has paid attention to serious conflict between the massive inaccuracies caused by personal differences and the required high accuracy and reliability for characterizing chronic diseases. The misuse of randomized clinical trials was mainly driven by a misplaced incentive to avoid so-called selection bias and quality of evidence. The breached core presumption (or lack of accuracy and reliability), plus improper averaging of positive-and-negative treatment’s effects, plus interference of multiple interfering factors, plus the stringent rejection criteria or low p values inevitably resulted in biased or wrong conclusions in past medical studies. The research model is unable to determine the benefits of any weak and slow treatment which could be vitally important for correcting subtle imbalances in the human body. The misuse of clinical trials are predictably responsible for the failure to find treatment effects for chronic diseases and failure to identify harmful effects

of toxic compounds in environment. In sum, clinical trail should be rejected because it offers no chance to find cures under any of our theoretical and practical models mimicking real diseases. Our findings may be similarly applicable to randomized controlled trials used in social sciences, environmental studies, life sciences, etc. as long as those required conditions are met.

ADDITIONAL INFORMATION

Supplementary information is provided.

FUNDING STATEMENT

The author(s) declared that no grants were involved in supporting this work.

REFERENCES

AccessMedicine Clinical Laboratory Reference Values https://accessmedicine.mhmedical.com/content.aspx?bookid=2503&sectionid=201361245 (accessed on July 2, 2019).

Albero A, Lopéz JE, Torres A, de la Cruz L, Martín T. Effectiveness of c hemotherapy in advanced differentiated thyroid cancer: a systematic review. Endocr Relat Cancer. 2016 Feb;23(2):R71-84. doi: 10.1530/ERC-15-0194.

Altman DG. Randomisation. BMJ 1991; 302: 1481-2.

Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer. 1954;8(1):1–12.

Ashdown ML, Robinson AP, Yatomi-Clarke SL, Ashdown ML, Allison A, Abbott D, Markovic SN, and Coventry BJ. Chemotherapy for Late-Stage Cancer Patients: Meta-Analysis of Complete Response Rates, Version 1. F1000Res. 2015; 4:232.

Booth FW, Roberts CK, and Laye MJ. Lack of exercise is a major cause of chronic diseases. Compr Physiol. 2012 Apr; 2(2): 1143–1211.

Brydon L, Magid K, Steptoe A. Platelets, coronary heart disease, and

stress. Brain Behav Immun (2006) 20: 113–119.

Campbell, SK. Flaws and Fallacies in Statistical Thinking. Prentice-Hall Inc, 1974.

Cheng L, Swartz MD, Zhao H, Kapadia AS, Lai D, P. Rowan J, Buchholz TA, Giordano SH. Hazard of recurrence among women after primary breast cancer treatment—A 10-year follow-up using data from SEER-medicare. Cancer Epidemiol. Biomarkers Prev. 2012; 21,800–809.

CHMP (Committee for Medicinal Products for Human Use). Guideline on adjustment for baseline covariates in clinical trials. 26 February 2015 EMA/CHMP/295050/2013.

Colleoni M, Sun Z, Price KN, Karlsson P, Forbes JF, Thürlimann B, Gianni L, Castiglione M, Gelber RD, Coates AS, Goldhirsch A. Annual hazard rates of recurrence for breast cancer during 24 years of follow-up: Results from the international breast cancer study group trials I to V. J. Clin. Oncol. 34, 927–935 (2016).

Cormie P, Zopf EM, Zhang X, Schmitz KH. The Impact of Exercise on Cancer Mortality, Recurrence, and Treatment-Related Adverse Effects. Epidemiologic Reviews, Volume 39, Issue 1, January 2017, Pages 71–92,

Deaton A, Cartwright N, Understanding and misunderstanding randomized controlled trials. Soc Sci Med. 2018 August; 210: 2–21.

Demicheli R, Retsky MW, W. Hrushesky JM, Baum M. Tumor dormancy and surgery-driven interruption of dormancy in breast cancer: Learning from failures. Nat. Clin. Pract. Oncol. 2007; 4,699–710.

Egbewale BE, Lewis M, and Simcorresponding J, Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. BMC Med Res Methodol. 2014; 14: 49.

Fleiss JL, Levin B, Park MC. A statistical Methods for Rates and Proportion. 3rd ed. Hoboken NJ: John Wiley and Sons; 2003. How to randomize.

Giovannucci E, Harlan DM, Archer MC, Bergenstal RM, Gapstur SM, Habel LA, Pollak M, Regensteiner JG, and Yee D. Diabetes Care. 2010 Jul; 33(7):1674–1685.

Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. Br Med J 1977;1:85-7.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN,

Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016; 31:337–350.

Kalish LA, Begg GB. Treatment allocation methods in clinical trials a review. Stat Med. 1985;4:129–44.

Hall JC, Hill D, Watts JM. Misuse of statistical methods in the Australasian surgical literature. Aust N Z J Surg 1982;52:541-3.

Kalben BB. Why Men Die Younger: Causes of Mortality Differences by Sex. North American Actuarial Journal. 2000;4:83–111.

Kim JS, Kim DK, and Hong SJ. Assessment of errors and misused statistics in dental research. Int Dent J 2011;61:163-7.

Krall JA, Ferenc R, Mercury OA, Pattabiraman DR, Brooks MW, Dougan M, Lambert AW, Bierie B, Ploegh HL, Dougan SK, Weinberg RA. The systemic response to surgery triggers the outgrowth of distant immune-controlled tumors in mouse models of dormancy. Sci. Transl. Med. 10, eaan3464 (2018).

Levine EM & Robins EB. Differential temperature sensitivity of normal and cancer cells in culture. Journal of Cellular Physiology. 1970;76(3):373–37 9·

Liao JB, Viruses and Human Cancer. Yale J Biol Med. 2006 Dec; 79(3-4): 115–122.

Maeda H and Khatami M, Analyses of repeated failures in cancer thera py for solid tumors: poor tumor-selective drug delivery, low therapeutic effic acy and unsustainable costs, Clin Transl Med. 2018; 7:11.

Millimouno FM, Dong J, Yang L, Li J, and Li X. Targeting Apoptosis Pathways in Cancer and Perspectives with Natural Compounds from Mother Nature. Cancer Prev Res; 2014; 7(11); 1081–107.

Milken Institute. Milken Institute Study: Chronic Disease Costs U.S. Economy More Than $1 Trillion Annually. https://www.fightchronicdisease.org/latest-news/milken-institute-study-chronic-disease-costs-us-economy-more-1-trillion-annually. Last assessed on July 2, 2019.

Navarro SL, Kantor ED, Song X, Milne GL, Lampe JW, Kratz M, White E. Factors associated with multiple biomarkers of systemic inflammation. Cancer Epidemiology Biomarkers and Prevention. 2016;25(3),521.

Nemeroff CB, Musselman DL, Evans DL. Depression and cardiac disease. Depress Anxiety 8 (Suppl 1): 1998;71–79.

Nowell, PC. The clonal evolution of tumor cell populations. Science. 1976;194 (4260): 23–28.

Ogino S, Fuchs CS, Giovannucci E (2012). "How many molecular subtypes? Implications of the unique tumor principle in personalized medicine". Expert Rev Mol Diagn. 12 (6): 621–8.

Ogino S, Lochhead P, Chan AT, Nishihara R, Cho E, Wolpin BM, Meyerhardt JA, Meissner A, Schernhammer ES, Fuchs CS, Giovannucci E (2013). "Molecular pathological epidemiology of epigenetics: Emerging integrative science to analyze environment, host, and disease". Mod Pathol. 26 (4): 465–84.

Prasad S, Sung B, and Aggarwal BB. Age-Associated Chronic Diseases Require Age-Old Medicine: Role of Chronic Inflammation. Prev Med. 2012 May; 54(Suppl): S29–S37.

Raghupathi W. and Raghupathi V. An Empirical Study of Chronic Diseases in the United States: A Visual Analytics Approach to Public Health. Int. J. Environ. Res. Public Health 2018, 15, 431.

Sagar SM, Yance D, and Wong RK. Natural health products that inhibit angiogenesis: a potential source for investigational new agents to treat cancer—Part 1. Curr Oncol. 2006 Feb; 13(1): 14–26.

Salani R, Backes FJ, Fung MF, et al. Posttreatment surveillance and diagnosis of recurrence in women with gynecologic malignancies: Society of Gynecologic Oncologists recommendations. Am J Obstet Gynecol. 2011;204(6):466–478.

Schuetz JM, Grundy A, Lee DG, Lai AS, Kobayashi LC, Richardson H, et al. Genetic variants in genes related to inflammation, apoptosis and autophagy in breast cancer risk. PLOS ONE 2019;14(1): e0209010. https://doi.org/10.1371/journal.pone.0209010

Schul KF, Grimes DA. Allocation concealment in randomized trials: Defending against deciphering. Lancet. 2002;359:614–8.

Segerstrom SC, Miller GE. Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry. Psychological Bulletin 2004;130(4):601-630.

Sloan EK, Priceman SJ, Cox BF, et al. The sympathetic nervous system induces a metastatic switch in primary breast cancer. Cancer Research 2010;70(18):7042-7052.

Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research – a review of common pitfalls, SWIS SMED WKLY 2007;

137:44–49.

Sundquist J, Li X, Johansson SE, Sundquist K. Depression as a predictor of hospitalization due to coronary heart disease. Am J Prev Med 2005; 29: 428–433.

Tinker, A. How to Improve Patient Outcomes for Chronic Diseases and Comorbidities. Available online: http://www.healthcatalyst.com/wp-content/uploads/2014/04/How-to-Improve-Patient-Outcomes.pdf Last assessed on July 2, 2019.

Ueshima H, Sekikawa A, Miura K, et al. Cardiovascular diseaseand risk factors in Asia: a selected review. Circulation 2008 Dec16;118(25):2709.

Vivarelli S, Salemi R, Candido S, Gut Microbiota and Cancer: From Pathogenesis to Therapy. Cancers (Basel). 2019 Jan; 11(1): 38.

White SJ. Statistical errors in papers in the British Journal of Psychiatry. Br J Psychiatry 1979;135:336-42.

Yeung PK, Wong JT. Inhibition of cell proliferation by mechanical agitat ion involves transient cell cycle arrest at G1 phase in dinoflagellates. Protopl asma. 2003 Mar;220(3-4):173-8.

Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet 2004;364; 937-52.

Supplement to: Randomized Controlled Clinical Trial Is Inherently Biased and Invalid In Studying Chronic Diseases, As Compared with Multiple Factors Optimization Trial

This appendix is provided by the authors to provide additional information and evidence for their study.

A. Changing Speeds of Health Properties in Chronic Conditions

The exemplar calculation shows how small changes cause chronic diseases.

(1) The glucose "normal range" is said to be 3.89-5.50 (6.10) mmol/L. In a hypothetical person, the optimum level of 4.0 mmol/L will not result in fat accumulation. Assuming that the glucose level is raised by 25% or 1.0 mmol/L, and only 1% (e.g., 0.01 mmol/L) of the extra glucose is deposited on the body, it can create serious consequence. The concentration of 1.0 mmol/L would be 0.001 mol/L×180 g/mol = 0.18 g/L. Each litter of blood contains additional 0.18 grams glucose. If the person has an average heart output of 6 litters per minute, the total heart output volume each year is 6×60×24×365 = 3,153,600 L. So, the total extra glucose that would be available for storage as fats is 3,153,600 L×0.18 g/L×1%=5.7 kg, which is equivalent to 5.7*(4/9)=2.5 kg each year.

(2) Capillaries, important components of micro-vascular network, are small blood vessels from 5 to 10 micrometers (μm) in the inner diameter. The capillary density in tissues and capillary inner diameters determine blood flow resistance in the segment. Flow resistance for any blood vessel segment can be computed by using $R=8\eta l/\pi r^4$, where, η is viscosity of blood, l is the length of blood vessel, and r is the inner radius of the blood vessel. Assuming that a capillary of 10 μm has been coated with 1 μm thickness fats in its inner wall, and a one-year exercise helps remove the deposited fats, the radium of each capillary is increased by (5-4)/4*100 = 25%. So, the exercise reduces the flow resistance of the capillary by 59%. The rate of removal is 1/365=0.0027 μm per day.

(3) A person with 10 cancer cells that grow at 0.1% (increase one net cell for one thousand cancer cells), the total cancer cell number is estimated to 32.4 billion after sixty years. A 10% increase in the rate constant from 0.01 to 0.011 for a tumor of 500 cells will increase the final cancer cell

number from 42 billion to 261 billion in five years. A 1% increase in the apparent rate constant, 0.01, will increase the final cancer cell number by a multiplying factor of 1.2 in five years (42.25 to 50.59 billion). Regardless of cancer causes and detailed mechanisms, cancer outcome depends on the imbalance between cancer cell death rate and cell division rate.

(4) Some human physiological properties must be maintained in narrow ranges. Normal body temperature is in a range from 97°F (36.1°C) to 99°F (37.2°C). The pH of the human blood is maintained in a tight range between 7.35 and 7.45, and any minor deviations from the personal optimal numbers can have health implications.

(5) In vertebral body replacement, shape and size of a placement vertebral body structure must match exactly the original one to be replaced. If the replacement part has one millimeter extra, it may cause great discomfort and pain. Denture must match mouth mounting member exactly. Structural imbalance can be found in joint diseases. A 1 mm outgrowth in bones in five years means very small change in a given time interval.

B. Clinical Trials Were Mainly Used in Studying Acute Health Problems in the Early History

The world's first clinical trial is recorded in the “Book of Daniel” in The Bible [Legumes, 2009]. In Ambroise Pare trial conduced in 1537, the purpose was to treat wounds of battlefield-wounded soldiers [Legumes, 2009]. Two hundreds of years later, James Lind (1716-94), the first physician, conducted a controlled clinical trial to treat scurvy, a vitamin C deficiency [Legumes, 2009; Twyman, 2004]. The word placebo first appeared in medical literature in the early 1800s [Legumes, 2009]. In 1863, U.S. physician Austin Flint planned the first clinical study, comparing a dummy remedy to an herbal extract for patients suffering from rheumatism. The Medical Research Council UK carried out a trial in 1943 to investigate patulin (Penicillium patulinuman extract) treatment for the common cold [Hart, 1999] and this study was controlled by keeping the physician and the patient blinded to the treatment. A first randomized control trial was carried out in 1946 by MRC of the United Kingdom for treatment of streptomycin in pulmonary tuberculosis [Hart 1999; MRC 1948]. In parallel to the development of clinical trials was evolution of ethical and regulatory framework, which shaped ethics of human experimentation and clinical practices. The FDA became a law enforcement organization after the US Congress passed the Food and Drugs Act in 1906 and regulate drug approvals for the U.S.

After randomized clinical trial is widely accepted, no study has been

done to study whether it is a competent approach for studying chronic diseases.

C. Lack of "Nearly Identical Units" in Clinical Trials

To explore the limitations of population-based clinical trials, we examine the machine-repairing model used in the auto industry. Auto mechanics always focus on structures and functions of individual cars, but never use information from other cars. This individualized approach is used in the entire machine industry, covering cars, TV sets, computers, airplanes, etc.

We establish two hypothetical repair models to explore whether a population-based repair model could work. In the first one, all cars made by Honda will be diagnosed and repaired by using the performance data which is acquired from all Honda cars such as Accord, Civic, Honda Fit, Honda CR-V, and Honda Pilot, etc. In this hypothetical model, even though most parts are similar in structure and function, they vary in size, shape and capacity. Most repair attempts would fail. If a lucky attempt makes a broken car to run, the car most probably cannot be restored to its optimum condition.

In the second hypothetical model, car performance and repairing data is acquired from all makes and models of cars in the world. Such population data is then used as guidance in repairing any car from any make. In this hypothetical model, the performance data acquired from all cars are summed and averaged across makes, models, mileages, mechanical conditions, accident histories, etc. We anticipate that few or no mechanical problems in cars can ever be fixed.

Even a moderately complex machine such as a car requires balance among individual components. Each component must be able to mount in an exact location, have a required installation space, have suitable structural strength, and optionally use a right amount of power or energy. In complex machines, all key components must maintain balances in fuel flow, heat exchange, lubricant usage, etc.

D. Large Personal Differences Implied by Exemplar Reference Ranges of Health Properties Established for the Human Population

We will show some of well known health properties in the table below. This reflects huge differences in the human population.

Table S1. Reference Ranges of Laboratory Tests

	Specimen	**SI Reference Interval**	**SI Units**
Alanine aminotransferase	Serum	10–40	U/L
Albumin	Serum	35–50	g/L
Aluminum	Serum, plasma	0.0–222.4	nmol/L
Alanine	Plasma	210–661	µmol/L
Ammonia (NH_3)	Plasma	11–35	µmol/L
β-Carotene	Serum	0.2–1.6	µmol/L
HDL (Adequate)	Plasma	1.03–1.55	mmol/L
LDL Near optimal	Plasma	2.59–3.34	mmol/L
LDLBorderline high	Plasma	3.37–4.12	mmol/L
LDL High	Plasma	4.15–4.90	mmol/L
Cholesterol (total)	Serum	1.3–5.20	mmol/L
Platelet count	Whole blood	150–450	10^9 L^{-1}
Red blood cell (Female)	Whole blood	3.9–5.5	10^{12} L^{-1}
Red blood cell (Male)	Whole blood	4.6–6.0	10^{12} L^{-1}
White blood cell count	Whole blood	4.5–11.0	10^9 L^{-1}

Fatty acids (nonesterified)	Plasma	0.28–0.89	mmol/L
Glucose	Serum, plasma	3.9–6.1	mmol/L
Triglycerides	Plasma, serum	0.11–2.15	mmol/L
Vitamin A (retinol)	Serum	1.05–2.80	µmol/L
Vitamin B1 (thiamine)	Whole blood	74–222	nmol/L
Vitamin B5 (pantothenic acid)	Whole blood	0.9–8.2	µmol/L
Vitamin B6 (pyridoxine)	Plasma	20–121	nmol/L
Vitamin B12 (cyanocobalamin)	Serum	118–701	pmol/L
Vitamin C (ascorbic acid)	Plasma, serum	23–85	µmol/L
Vitamin D, 1,25-dihydroxyvitamin D	Plasma, serum	42–169	pmol/L
Vitamin E (α-tocopherol)	Plasma, serum	12–42	µmol/L
Vitamin K	Plasma, serum	0.29–2.64	nmol/L

Source: AccessMedicine [AccessMedicine].

E. Relatively Large Effects of Causal and Interfering Factors In Clinical Trials and Their Impacts on the Error Term

We have noted that the requirement of using “nearly identical units” in a clinical trial is most probably violated if the trial is used to study chronic

diseases. The trial has one treatment with an effect of μ (the first treatment factor u could be a random variable too). Each observed value of the health property to be measured in a trial may be expressed in following equation.

$$x_{ij}=\mu+\delta_j+\varepsilon_{ij,} \quad (1)$$

where j=1, 2,…,+s (the number of treatments)

i=1, 2,…,+n (the number of data per treatment level)

$$\text{where } a\varepsilon_{ij}=s\varepsilon_{ij}+\varepsilon_{ij,} \quad (2)$$

$$\text{where } s\varepsilon_{ij}=s_{1ij}+s_{2ij}+\ldots., s_{knm}. \quad (3)$$

δ_j is the effect of treatment's level j, ε_{ij} is the uncontrollable random errors which must be much smaller than δ_j, $s\varepsilon_{ij}$ is the effect caused by a series of interfering factors, and $s\varepsilon_{ij}$ may be viewed as part of the apparent error because each data point is be affected by ($s_{1..}+s_{2..}+\ldots., s_{k..}$). The total number of interfering factors may be several, tens to even hundreds. Those factors affect every data point under any types of experimental designs. $a\varepsilon_{ij}$ is the apparent error that is actually measured or detected in the trial. An implied presumption is that $a\varepsilon_{ij}$ must be much smaller than $\mu+\delta_j$. A trial conclusion may be still useful if the apparent error $a\varepsilon_{ij}$ (which includes $\varepsilon_{ij,}$ and $s_{1..}+s_{2..}+\ldots., s_{k..}$) is still much smaller than $\mu+\delta_j$ so that the total interfering effects can be neglected in practice.

In traditional clinical trials, most $s\varepsilon_{ij}$ terms were not be identified, and nor controlled, they are simply added to the apparent error term $a\varepsilon_{ij}$. Assuming that those uncontrolled factors follow normal distributions with respective parameters:

$$\varepsilon_{ij} \sim N(0, \sigma_E^2)$$

$$s_{1ij} \sim N(\mu_1, \sigma_1^2)$$

$$s_{2ij} \sim N(\mu_2, \sigma_2^2)$$

….

$$s_{kij} \sim N(\mu_k, \sigma_k^2)$$

The apparent error term is the sum of the error ε_{ij}, plus all uncontrolled interfering factors ($s_{1ij}+s_{2ij}+\ldots., s_{kij}$), plus interaction terms, (which are omitted to make the model simpler). Assuming that all uncontrolled factors are independent, the apparent error $a\varepsilon_{ij}$ also follows a normal distribution [Wikipedia (2)]:

$$a\varepsilon_{ij} \sim N(\mu_t, \sigma_t^2)$$

$$\mu_t = \mu_1 + \mu_2 + \ldots, + \mu_k \quad (4)$$

$$\sigma_t^2 = \sigma_E^2 + \sigma_1^2 + \sigma_2^2 + , \ldots, \sigma_k^2 \quad (5)$$

The apparent error term is routinely used in statistical analysis. It size must be considered in a trial directed to evaluating weak treatment. This implies that if all interfering factors are used as part of the treatment, it can raise the treatment effect and reduce the variances of the error term.

Proof can be made by taking any two random variables, per known theorems or the known methods such as "the Sum of normally distributed random variables" [Wikipedia (1)]. The sum, which is also a random variable, has increased mean and combined variances. This summed random variable is further added to a third random variable. The trend of increasing variances for the sum of a limited number of random variables, s_1 to s_k, can be generalized.

All interfering factors are random variables with probability density functions $f(s_1)$, $f(s_2)$,...., $f(s_k)$. Therefore, the distribution of the sum of those factors could be created by drawing: when the error takes a particular value, $s_{1..}$ can take any of its possible values according to its distribution probability density. Thus, the random variable $s_{1..}$ is added to the error term ε_{ij} to generate a new random variable. By repeating this process, all random variables from s_{2ij} to s_{kij} are added into the error term to become a final apparent error. By repeating this process, one could get an empirical distribution data of the sum of all random variables. By using computer, one could create an apparent error distribution for any interfering factors that follow other types of distribution.

F. Hypothesis Test for Comparing Two Populations' Means

In a typical clinical trial, the purpose is to determine if a treatment is different from a control, the trial results in two sets of measures X=X1, X2...., Xn (for the treatment) and Y=Y1, Y2...., Yn (for a control). We assigned a start patient survival data in days in Table S2 below, and assumed that the treatment can be adjusted by strengthening or weakening its treatment effects, we will get following data sets.

Table S2. A Hypothetical Test Data Using Two Population Means

Ctrl Srvl.	Yi-Y	$(Yi-Y)^2$	True	TX Srvl.	Xi-X	$(Xi-X)^2$

(days)			Effect (days)	(days)		
130	-75	5625	57	187	-75	5625
160	-45	2025	57	217	-45	2025
190	-15	225	57	247	-15	225
220	15	225	57	277	15	225
250	-45	2025	57	307	45	2025
280	75	5625	57	337	75	5625

From the hypothetical data, we got following statistical parameters:

For the Control: n1 is the control sample number, Y is the survival mean for the control, and Sy^2 is the mean squares of the control.

For the Treatment: n2 is treatment sample number, X is the survival mean for the treatment, and S_x^2 is the mean squares of the treatment. Assuming that all model conditions are met (which is not possible due to the nature of this simulation), and that the variances in the control and treatment are consistent. We conduct the hypothesis test below:

$$Sy^2=\frac{1}{n1-1}\sum(Yi-\overline{Y})^2$$

$$Sx^2=\frac{1}{n2-1}\sum(Xi-\overline{X})^2 \quad (6)$$

$$Sw^2=\frac{(n1-1)Sy^2+(n2-1)Sx^2}{n1+n2-2} \quad (7)$$

$t_{0.05}(n1+n2-2)$ is found from a t-table.

$$If\ \overline{X}-\overline{Y}\geqslant t_{0.05}(10)Sw*\sqrt{\frac{1}{n1}+\frac{1}{n2}}, \quad (8)$$

reject H_0.

We got following statistical parameters:

For the control: $|Y|=205$, $s_y^2=3150$

For the treatment: $|X|=262$, $s_x^2=3150$

$$Sw^2 = \frac{(6-1)3150 + (6-1)3150}{6+6-2} = 3150$$

Find t value at p=0.05: $t_{.05}(10)=1.81$.

$$t_{0.05}(10)\,Sw * \sqrt{\frac{1}{n1}+\frac{1}{n2}} = 1.81 * \sqrt{3150}\ \sqrt{\frac{1}{n1}+\frac{1}{n2}} = 58.65.$$

Since X–Y =57 < 58.7, accept the null hypothesis. This outcome is against the assumed fact that the treatment can actually extend survival by 57 days. In conducting the hypothesis test, the final outcome is determined by comparing the treatment benefit X–Y with mean squares (S_x^2, S_y^2) or standard error attributable to personal differences. When survival times are widely dispersed among patients, the treatment's effect is hidden in the experimental error.

The mathematical operations reveal that the hypothesis test outcome depends on treatment's mean, X–Y, and S_y^2 and S_x^2. A spreadsheet data set can be constructed, which allows for changing the data in the first column so that one can see how hypothesis test outcomes would change with data being manipulated.

(1) If data dispersion is fixed, the chance of rejection entirely depends on the effect of the treatment. When S_y^2=3150, S_x^2=3150, and S_w^2=3150 are held constant, the point of rejection is a constant. When the treatment's effect is increased to 59 days or any reasonable days, the true treatment effect is confirmed at p=0.05.

(2) If data dispersion is held as zero, survival times become ordinary variable. Zero variances may be viewed as the limit of reducing variances. To avoid the density function vanishes, several hour times are added as noise to the control data set so that S_x^2=0.035, S_y^2=0.035, and S_w^2=0.035. In this case, even one day extension of survival times can be confirmed at p=0.05 (Ignoring practical difficulty in setting up H_0 and H_1). By changing survival times, the following results were obtained.

Table S3. Rejection Value (X-Y) for the Same Probability Increases with Control's and Treatment's Variances.

S_y^2 for Control	0.035	350	3150	12655	35000

S_x^2 for Treatment	0.035	350	3150	12655	35000
S_w^2	0.035	350	3150	12655	35000
Rejection points (at p=0.05)	>0.19	>19.6	>58.7	>117.3	>195.5
Min X-Y for rejecting H_0	1	20	59	118	196

The above table shows that when data dispersion increases, the rejection point at the same probability dramatically increases. However, if a patient population is selected with great differences in their baseline survival times, even 195 days survival time extension were not recognized due to the type II error (false acceptance of the null hypothesis). This problem is well known in statistics, but what is shown is that in most, if not all, clinical trials, expected variances are sufficiently large to result in consistent failure to recognize weak and slow treatment effects. Here, the survival times are reasonable numbers found in cancer literature.

S_w^2 can be raised by casual and interfering factors showed in Model A in the Method Section.

(3) When sample sizes (patient numbers) in the control and the treatment are sufficiently large, the acceptance region for H_o is determined by the following range:

$$\text{X-Y} \pm Z_{0.05} * \sqrt{\frac{S_y^2}{n1} + \frac{S_x^2}{n2}}$$

In this case, the rejection point is determined by using the normal distribution rather than the t-student distribution.

(4) If the treatment has the effect of extending more survival time for each of the patients, it will result in a larger X–Y, which is directly compared with a value defining the rejection region. While multiple factors bundled into the treatment may increase data dispersion of the control and the treatment, it tends to move into the region for rejecting the null hypothesis faster, resulting in recognizing overall effects of the treatment.

G. Hypothesis Test For Comparing Paired Differences

In the following hypothetical test, we show that variables controlled trials tend to fail to recognize single weak and slow factor, as a result of the acceptance of the null hypothesis.

There are N persons with a health property x being observed before a treatment and after a treatment. It is assumed that the health property before the treatment and after the treatment can be accurately measured. A treatment may comprise one treatment component or factor F selected from F1, F2…., Fn. For all patients, the trial would result in a series of paired data: X1=x'1-x1, X2=x'2-x2,…., Xn=x'n-xn, where x' is a health property after a treatment, x is the value of the property before the treatment, and Xi means their difference.

In this test model, a systolic blood pressure is used as the health property. The treatment is a weak single factor, which can alter blood pressure by only 1.5 mm Hg. We first tried six data points with blood pressure range from 145 to 180 mm Hg, and then added some random noises to the data in an arbitrary way. We want to see whether the true effect of the treatment could be confirmed in the hypothesis test. We generated following data:

Table S4. Blood Pressure Data In a Hypothetical Trial

Assumed Sys. BP mm Hg	Treatment Real Effect (mg Hg)	Fluctuations (mm Hg)	Predicted Change Xi=(y'i-yi)	Mean changes X	$(Xi-X)^2$
161	-1.5	2	+0.5	(-1.5)	4
180	-1.5	-2	-3.5	same	4
130	-1.5	2	+0.5	same	4
150	-1.5	-2	-3.5	same	4
145	-1.5	2	+0.5	same	4
179	-1.5	-2	-3.5	same	4

By following statistical steps, it is assumed, obviously against assumed treatment effects, that the treatment had no real effect and all changes in measurements were caused by random error. The data would follow a distribution centered at zero. So, the task is to determine if y'i-yi belongs to the normal distribution, $N(0, \sigma^2)$. The test starts with setting a null

hypothesis: E(X)=0, with alternative hypothesis being, E(X)<0:

$$|\overline{X}|=\frac{1}{n}\sum Xi; \quad (9)$$

$$S^2=\frac{\sum (Xi-\overline{X})^2}{n-1} \quad (10)$$

$$If\ |\overline{X}|>t_a\frac{s}{\sqrt{n}}, \quad (11)$$

reject the null hypothesis.

From the data, one can find s^2= 4.9, SD=2.19, and find from t-distribution table, $t_{0.05}(5)$=2.01.

$$t_a\frac{s}{\sqrt{n}}=1.79$$

Since the mean |X|= |-1.5| < 1.79, the hypothesis test accepts the null hypothesis. The finding that the treatment is ineffective is contrary to the presumed fact that the treatment has 1.5 mm Hg reduction. Here, the treatment is a weak treatment. This outcome implies, as expected, that when error attributable to measurements is larger than the treatment effect, such a small effect is not recognized.

We set up a spreadsheet data set with variables that can be changed. We could repeat the same simulations by using a much larger data set. Measurement error may be much larger than 2 mm Hg, but this does not change the general pattern that weak treatment will not be recognized due to type II error.

S^2 can be raised by all casual and interfering factors showed in Model A in the Method Section. Those causal and interfering factors make patients to respond to the same treatment in more different ways.

Assuming that the same treatment is optimized by using several factors to treat blood pressure, and the treatment could contain following components:

(1) Jog one hour each morning, which is assumed to generate an effect of lowering 10 mm Hg by removing fats from inner walls of blood vessels.

(2) Administrate a heavy metal deduction program, which is assumed to result in a 5 mm Hg reduction by reducing damages to blood vessels.

(3) Practice meditation daily to help blood vessels to achieve relaxed state. It is assumed to reduce blood pressure by 5 mm Hg.

(4) Reduce and avoid refined foods, fast foods, fried foods, etc. for one year. It is assumed to reduce the blood pressure by 5 mm Hg.

(5) Correct vitamin deficiency to improve the brain's regulatory function which is assumed to lower blood pressure by 5 mm Hg.

(6) Reduce life stress, job stress, and emotional stress, etc. to improve hormonal regulations, which is assumed to reduce blood pressure by 5 mm Hg.

(7) Improve the kidney functions to improve the efficiency of removing metabolic toxic by-products. It is assumed to reduce blood pressure by 5 mm Hg.

(8) Adjust fat compositions for omega 6/3 fatty acids ratio to a normal range in diet, which is assumed to lower blood pressure by 5 mm Hg.

It is further assumed that those weak effects work slowly. If the trial is not sufficiently longer, the treatment factors may deliver only part of their respective maximum effects.

The above factors may interact with each other. If blood vessels are enlarged, the brain's regulation of the vascular system is improved, damages to blood vessels are cured, toxic compounds are removed, and inflammation is reduced, total blood pressure reduction will be more than the sum of all assumed individual effects. If similar simulations are conducted by using various combination factors, the chances of rejecting the null hypothesis rapidly increase, thereby affirming the treatment's true heath benefits.

If more factors are included in a treatment, the data set will come out with a practical effect of increasing the likelihood of rejecting the null hypothesis. This is because that $|X|$ is increased while s^2 is reduced by making all factors work in a similar way on all patients. This implies that optimization using as many factors can yield a result of the treatment. If the treatment comprises factors 1 and 2, the test could result in $|X|=15>1.79$. If the treatment comprises factors 2, 3, 4, 5, the test would result in 20 mm Hg reduction. If all factors are used, the treatment might reach 45 mm Hg as the potential maximum.

Looking at the logic, the variances are caused by $(Xi-X)^2$. If the treatment has a total net effect, the variances depend on how the treatment effects are dispersed among individual patients. If all patients are very consistent, and their net treatment effects are close to the mean, the test would be able to recognize smaller treatment effects. If some patients show big treatment effects, but others show small effects, the large differences will result in large variances and the value for defining the rejection region for rejecting the null hypothesis will increase per the equation (11).

Whether a true treatment effect can be detected by a hypothesis test depends on whether all patients respond to the treatment in a similar way. To have the true treatment effect determined accurately, a basic requirement is that all patients will respond to the treatment in a quantitatively similar way. In reality, a clinical trial must introduce massive variances attributable to personal differences in genetics, phenotypes and mental conditions. If 3 out of 6 patients are cured, there is no point to use the three poor outcomes to refute the treatment benefits. There is no justification to use response dispersion as a basis to refute treatment benefits for individual patients.

This example is equivalent to Model B in the Method Section. When multiple factors are bundled together, the total treatment effects are much larger than the experimental error. This example also shows that cures for chronic diseases lie in optimizing as many factors as possible to achieve best curative results.

H. Effects of Interfering Factors on Variance Analysis

We will review basic assumptions used in variance analysis and then will evaluate the presumptions when the clinical trial is used to study chronic diseases. Variance analysis is based on the following basic model:

$x_{ij} = \mu + \delta_j + \varepsilon_{ij}$ i=1, 2,...., n (Sample No. within a treatment)

$\varepsilon_{ij} \sim N(0, \sigma^2)$ j=1, 2,...., s (Treatment No)

In this statistical model, statistical parameters, total sum of squares, error sum of squares, and treatment sum of squares can be determined by using following equations [Roussas, 1997]:

$$\bar{x}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{.j} \tag{12}$$

$$SS_E = \sum_{j=1}^{s} \sum_{i=1}^{nj} \left(x_{ij} - \bar{x}_{.j} \right)^2 \tag{13}$$

$$SS_A = \sum_{j=1}^{s} \left(x_{.j} - \bar{x} \right)^2 \tag{14}$$

$$SS_T = SS_A + SS_E \tag{15}$$

$$\frac{(n-s)SS_A}{(n-1)SS_E} \sim F(s-1, n-s)$$

$$If \frac{(n-s)SS_A}{(n-1)SS_E} > F_\alpha(s-1, n-s), \text{ reject } H_0 \qquad (16)$$

Whether a data set comes out with its F-statistic falling within the acceptance or rejection region depends on the ratio of S^2_A to S^2_E. If variances between different treatments are large while the variances between individual data units are small, a large F-statistic will result. This will more likely reject the null hypothesis at a preset probability value.

Per the model assumption, every observed experimental value must be the sum of its population mean, plus treatment effect, plus random error. The true error term $\varepsilon_{ij,}$ must be attributed to a random error, which cannot be controlled due to natural variations in the process. It is typically assumed to be normally distributed with zero mean and constant variances. Some classical ways of generating random errors are throwing a die, flipping a coin, using an identical machine to produce products, production yields produced by same production line, melting points of same material, etc. The classical trials show that truly random errors cannot be controlled. S^2_E cannot contain causal and interfering effects unless they are small enough to be ignored for convenience.

While ANOVA model does not directly impose any requirements on the quantitative values of S^2_A and S^2_E, there is an implied presumption that experimental error must be much smaller that the treatment's effect to make a study result useful. Moreover, the F-test used for ANOVA analysis has additional assumptions and limitations. In practice, small effects of uncontrolled factors may be merged into ε_{ij}, only if the total effect of those factors is sufficiently smaller than the treatment effect $\mu+\delta j$. If the treatment effect is close to experimental error ε_{ij}, F-statistic will be smaller which is compared with S^2_A to S^2_E ratio, thereby causing the data to move toward the acceptance region of the null hypothesis. Only if $\mu+\delta j$ is much larger than ε_{ij}, can an F-test have a practical meaning.

When clinical trials are used to study drug or treatment effects, ε_{ij} comprises contributions of a large number of other interfering factors. In variance analysis in medical research, the apparent error terms ε_{ij} can be divided into two terms in practice:

The apparent error, $a\varepsilon_{ij}=s\varepsilon_{ij}+\varepsilon_{ij,}$

where ε_{ij} is the true random error that cannot be controlled.

$s\varepsilon_{ij} = s_{1ij}+s_{2ij}+\ldots., s_{kij},$

where k is the total number of uncontrolled interfering factors.

Uncontrolled factors directly raise the error term's mean and its variances. When an influencing factor is randomized, what can be achieved is that the factor will have similar effects on all treatment groups and the control because the factor affects all data points in all groups by a similar probability. However, randomization cannot hold down the error term's variances, and nor its means. This can be easily seen by imagining how exercise, diet patter, emotional adjustment, toxic compound levels, etc. occur randomly among different patients. All patients in a treatment and a control do exercise at will. Some do a lot, some do little, and some do nothing. Since exercise has great impacts on cancer survival times, exercise along can make survival times in each group widely dispersed.

A condition for using clinical trials is that the total effect of the true random errors and all uncontrolled interfering factors is much smaller than test treatment's effect. This condition can be satisfied in trials involving acute diseases. This condition is essentially always breached in trials that are used to study chronic diseases. Even though those factors will not affect the differences between the treatment and the control, they raise the error term's variances. A massively increased σ_t^2 still make trial outcomes meaningless. Since the effect of the treatment is small, the only way to improve treatment effect is using multiple treatment factors.

I. Simulation Shows How Uncontrolled Interfering Factors Distort Test Outcomes in F Tests

In the next example, we will show how separating some interfering factors in a one-factor variance analysis will change the hypothesis test outcome. We create data for one factor variance analysis (when treatment levels B1, B2, B3, and B4 are ignored as if they did not exist).

Table S5. Hypothetical Cancer Survival Data for a Treatment Factor and Some Unidentified and Uncontrolled Causal and Interfering Factors.

	A1	A2	A3
(B1)	100	320	530
(B2)	500	740	970
(B3)	900	1160	1480

(B4)	1400	1700	1950

In a first hypothetical case, B factors were not identified, thus, the trial was designed as a one-factor variance analysis. The results are:

SS_T=3,625,091

SS_A=515,117, S^2_A=257,558 (df=2)

$SS_E=S_T-S_A$=3,625,091-515,117=3,109,974

SS_E=3,109,974, S^2_E=345,552 (df=9)

Since $F_A=S^2_A/S^2_E$=257,558/345,552=0.75<$F(2, 9)_{0.05}$=4.26, accept H_0.

This case is similar to Model A where many unidentified causal and interfering factors raise the experimental error S^2_E.

In a second case below, both A and B factors are identified, assuming that treatment B is the sum of the effects of all unidentified and uncontrolled interfering factors such as genetic composition, age, sex, diet, exercise, stress level, lifestyle, emotional condition, chronic stress, etc. Now, the trial is a two-factor design. The new results are:

SS_T=3,625,091

SS_A=515,116, S^2_A=257,558 (df=2)

SS_B=3,101,691, S^2_B=1,033,897 (df=3)

$SS_E=SS_T-SS_A-SS_B$=3625,091-515,116-3101,691=8284

S_E (df=6)=8,284/6=1381

Since $F_A=S^2_A/S^2_E$=257,558/1381= 187>$F(2, 6)_{0.05}$=5.14, reject H_0 for factor A.

Since $F_B=S^2_B/S^2_E$=1,033,897/1381=749>$F(3,6)_{0.05}$=4.76, reject H_0 for factor B.

SS_E in the first trial is the sum of SS_E and SS_B in the second trial (3,109,974=8,284+3,101,691). When the error term contains variances of random and uncontrolled errors and variances of other causal or interfering factors, the true effects of A1, A2, and A3 on survival times are not confirmed.

When uncontrolled factors are not addressed, they are merged into the experimental error term and raise the means and variances of the error. A

weak treatment effect cannot be determined due to inflated the experimental error. Root cause can be traced to personal deviations in clinical trials, and statistical analysis makes the problem worse by rejecting whatever effect which is close to the apparent experimental error. Indeed, one could see from the raw data that treatments A1, A2, and A3 have clear treatment effects.

While the simple data set is used for illustration purposes, the same conclusion could be seen from the computation steps.

J. χ2 Goodness-of-fit Test

In one sample test for a discrete outcome, hypotheses are set up against an appropriate comparator. The test relies on χ2 (chi-square) distribution which ranges from 0 to ∞.

1. The test details

The test selects a sample and computes descriptive statistics on the sample data, compute the sample size (n) and the proportions of participants in each response category (p1, p2,…., pk) where k represents the number of response categories, and finally determine the appropriate test statistic for the hypothesis test.

$$\chi = \sum \left(\frac{(O-E)^2}{E} \right)$$

In the test statistic, O=observed frequency and E=expected frequency in each of the response categories. The observed frequencies are those observed in the sample and the expected frequencies are computed. When conducting a χ2 test, the observed frequencies in each response category are compared with the frequencies that are expected if the null hypothesis were true. These expected frequencies are determined by allocating the sample to the response categories according to the distribution specified in H_0. This is done by multiplying the total observed sample size (n) by the proportions specified in the null hypothesis (p10, p20,..., pk0).

To ensure that the sample size is large enough for the use of the test statistic above, the sample size meets the following condition: min(np10, np20,..., npk0)>5. The formula for the test statistic is given below. Test statistic for testing H0: p1=p10, p2=p20,..., pk=$p_{k0.}$ The critical value in a table of probabilities for the chi-square distribution with degrees of freedom (df)=k-1.

This goodness-of-fit test is based on an implied presumption that all differences between the observed frequencies and expected frequencies is due to uncontrollable sampling error. However, the outcome of each patient in clinical trials for studying chronic diseases actually depends on random errors and effects of many interfering factors. Thus the outcome in each category is distorted by the factors. For example, an unknown factor makes some patients to appear in a particular category. A factor causes N patients to move from p1 to p2, and causes M patients to move from p2 to p1. When N and M are of the same value or close, their effects happen to cancel out. All of the effects of interfering factors are not reflected in the test statistic. Thus, the final test outcome depends only on sampling frequencies, but has nothing to do with H_o and H_1 hypotheses.

K. Common Frequency Tests

One type of test often used in biological science is to test the frequencies of certain events against expected frequencies.

1. The test details

A randomized trial is conducted to evaluate the effectiveness of a new pain killer as compared with old pain killer. The trial comprises a total of 100 patients. The outcome as follows:

H_0: p1=p2, H_1: p1≠p2 at α=0.05.

Treatment Group	Sample Size (n)	Number of Patients With Improved Condition	Proportions
New drug	50	23	0.46
Old drug	50	10	0.20

The sample size is adequate. There should be at least 5 successes and 5 failures in each comparison group: min(n1p1, n1(1-p1), n2p2, n2(1-p2))≥5.

$$z=\frac{p1-p2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n1}+\frac{1}{n2}\right)}}$$

$$\hat{p}=\frac{x1+x2}{n1+n2}$$

This test is based on an implied presumption that all differences between the observed frequencies and expected frequencies are caused by uncontrollable sampling error.

The outcome of each patient in a typical clinical trial involving a chronic disease actually depends on random errors plus a large number of factors that affect the development and reversal of the disease. Thus, the observed frequency in the disease category is affected by those uncontrolled factors. The above model does not reflect the complexity of disease process. Even if a final test outcome happened to be right, it has nothing to do H_o and H_1 hypotheses. Improved conditions could be caused by other causal and interfering factors.

L. Randomization Cannot Cure the Flaw Caused by Personal Differences

Randomizing human subjects can reduce the different impacts of interfering factors on the treatment and the control. All interfering factors may raise treatment's mean and the control's means by a similar amount. However, randomization cannot eliminate the effects on the error's means and error variances.

The benefits of randomization have been known for a long time. It is intended to avoid systematic bias as high age can influence surgical outcome [Kalish and Begg, 1985; Fleiss et al, 2003], and prevents selection bias researchers and patients from knowing to which group the subject will be assigned [Schul and Grimes, 2002]. All interfering factors, whether known or unknown, that may affect the outcomes can be similarly distributed among groups. This similarity is very important and allows for statistical inferences on the treatment effects. Also, it ensures that other factors except treatment do not affect the outcome. If the outcomes of the treatment group and control group show differences, this will be the only difference between the groups, leading to the conclusion that the difference is treatment induced [Altman, 1991].

The above analysis is correct only if the implied presumption is held: the treatment's effect is much larger than the effect of all sources of uncontrollable errors which include causal factors and interfering factors. This presumption fails to hold in a trial where the effect of the treatment is close to or even smaller than the total effect of the errors and interfering factors. If a statistical analysis can fix such a fundamental problem, developments in detection and separation technology would be unnecessary.

Interfering factors can raise the error term's variances and means and

thus cause the trials to breach the implied presumption used in statistical analysis. The null and alternative hypotheses are remote from the reality of the trial and the conclusion will be wrong except by accident [García-Pérez, 2012]. In other statistical models such as $\chi 2$ goodness-of-fit test and frequency test, where the models take account only drawing errors, any hypothetical test outcome does not reflect reality.

M. Stratification Cannot Correct the Clinical Trial's Bias

We can show that stratification cannot remedy the increased variances of the apparent error by interfering factors. One can see from the following diagram:

Table S6 Stratification for four groups of patients, each at 50%.

Data points	Take 50%	Results
A1, A2, A3, A4, A5, A6	3	A1, A3, A6
B1, B2, B3, B4, B5, B6, B7, B8	4	B2, B4, B7, B8,
C1, C2, C3, C4	2	C2, C3,
D1, D2, D3, D4, D5, D6, D7, D8 D9, D10, D11, D12, D13, D14	7	D1, D4, D6, D7, D9, D11, D13, D14

The issue we focus is how the differences among individual patients might have contributed to the outcome of a clinical trial when the treatment effect is weak. The causal and interfering factors can randomly affect individual data. The variances from the first strata (A1, A3, A6), and all others still exist. The only impacts are due to changed sample size and reduced degrees of freedom.

If measurements within strata have lower standard deviation, stratification gives smaller error estimation. It increases representation for groups within the population and reduce the chance of imbalanced baseline.

Studies addressing covariates make an implied assumption that all trial subjects are substantially similar, and certain factors such as sex, age, trial sites, diseases condition, etc. affect the baseline of the health properties to be measured [Wang et al. 2019]. Thus, different sex ratios in the treatment arm and the control arm could result in unbalanced baselines. Stratification could eliminate such baseline imbalance. However, stratification cannot address the averaging effect of positive and negative responses within each arm, enlarged variances of the apparent experimental error by co-causal and interfering factors. It actually reduces baseline imbalance at the cost of increasing the apparent error variance.

N. Personal Differences Are the Main Cause of Simpson's Paradox

Simpson's paradox (Simpson's reversal, Yule–Simpson effect, amalgamation paradox, or reversal paradox) is well known for quantitative data: a positive trend appears for two separate groups, whereas a negative trend appears when the groups are combined [Wagner, 1982].

This result is often encountered in medical research statistics [Wagner, 1982; Holt 2016; Franks et al., 2017]. It was believed that the paradox can be resolved when causal relations are appropriately taken care of in the statistical modeling. Although past focus was on the differences between groups, the real cause is actually variances from individual persons. In studying weak effects, each person must be presumed to be different from another person. If the same trial is repeated N times by using the exactly same subject, Simpson's paradox will not seen. The regression pattern or data trend from a single person must be unique, given massive differences in personal genetics, phenotype and emotional states. If individual person's data could be acquired, the data should have very small dispersion. When regression is conducted by using people from different subgroups, the subgroups may show different patterns. When their data are pooled, a different patter is seen. A striking example of subgroup difference is heart diseases between Asian people and Western people.

Regression analysis for weak and slow treatment is actually an attempt to build a trend across massive differences among individual persons. This data may be useful in social sciences, they have little utility as far as cures are concerned. The root cause is large variances at personal levels. The regression curve built on a large population is not applicable to any individual person except by accident.

O. Clinical Trail Lowers Treatment Benefits by Improper Averaging Effects While Optimization Trial Enhances Treatment Benefits

We will construct a model which mimics a typical clinical trail to show another fatal flaw. We then compared it with an optimization trial. Assuming that a treatment has both beneficial effects, neutral effects, or adverse or negating effects on different patients, we will determine how a clinical trial performs, as compared with an optimization trial in personalized medicine.

Table S7 Indiscriminate Application of Treatments in Clinical Trail Degrades "Statistically Detected" Treatment Effects While An Optimization Trial Enhances Treatment Effect (Based on Hypothetical Data)

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Col. 7

TXs or Cause Name	Assumed Ben. Resp. Rt. (%)	Non-Resp. Rt. (%)	Neg. Resp. Rt. (%)	Stat. Ben. Resp. Rt. X (%)	Opt. Trial Tx Comb.	Opt. Trial Ben. Resp. Rt. (%)
A	20	65	15	5	A+other	>20
B	10	84	6	4	B+other	>10
C	10	82	8	[2]	C+other	>10
D	5	91	4	[1]	D+other	>5
E	5	85	10	(-5)	E+other	>5
F	3	94	3	[0]	F+other	>3
G	2	86	12	(-10)	G+other	>2
Overall	55		[58]			≥55

It is assumed that frequency used in the table is a kind of properties that can be used for statistical analysis of health properties.

In the table, columns 2-5 are for clinical trials and columns 6-7 are for optimization trials. It is assumed that a disease is caused by seven causes A, B….G under column 1, and each treatment can correct one of seven causes. For convenience, each treatment is referred to by its correspondent cause (e.g., A, B,….,G). It is further assumed that each treatment has true benefits on some patients, no effects on some patients, and negating effects on some patients, as shown in columns 2 to 4.

In a randomized clinical trial, a treatment under test is indiscriminately applied to all patients in the treatment group and their "statistically determined" treatment effect is shown in column 5. When the treatment is used on a patient whose cause does not match the treatment, the treatment is assumed to cause adverse side-effects if the patient is unable to tolerate. A misapplied treatment could turn an existing balance into an imbalance. Inadvertent side effects are ignored. A well matched treatment does not cause negating effects. Thus, statistically detected treatment effects shown in column 5 are much lower than the assumed or true beneficial response rates (column 2). The statistically determined benefits of each treatment are due to the averaging effect of the treatment on all patients. If a treatment extends lives for some patients but shortens lives by the same amount for the same number of patients, the statistical mean for the treatment is nearly zero.

Clinical trails tend to underestimate treatment benefits for the chronic disease. When no averaging effects exit, all treatments A to G would able to

cure 55% patients if each patient tries each of A to G treatments in turn, provided that each treatment does not cause inadvertent side effects. However, in reality all treatments have side effects on some patients. The need to avoid side effects will limit how many treatments a patient can try. In reality, patients cannot try all available treatments one by one due to limited trial time, resources and the need to avoid risks. Under the current medical models, doctors are generally unable to select treatments according to matched causes for patients. The best bet is thus using treatments with highest response rates.

Influenced by the clinical trail approach, treatments C to G will not be approved for use or not offered as a first line treatment. When those treatments are evaluated in clinical trials, they are indiscriminately applied to all patients. Since they address rare causes, they can result in higher adverse response rates and negating response rates than beneficial responses rates shown in column 5. Moreover, true response rates for those treatments cannot be correctly detected in the trial due to interfering effects of other factors. In addition, even if the beneficial effects of treatment E can be found, its use cannot be justified. Under the current treatment model, patients are not treated according to their specific causes. If treatments like E and G are randomly applied to patients, they could result in higher adverse response rates.

To avoid excessive risks of exposure, treatments C to G may probably not be approved for commercial use or not recommended for use by doctors. Only treatments A and B are available as the first line drugs. Only “majority patients” whose disease causes are most popular in the population have available treatments. “Minority patients” whose diseases are caused by rare causes are out of luck. They always fall in non-response groups no matter which treatments they try. Thus, medicine will be able to deliver response rate of 9% in this case even though the treatments would treat at least 55% by assumption.

Negating effects can be justified by using the balance theory for human health. Human health is maintained by many balances such as calorie balance, nutritional balance, bone formation and resorption balance, pH balances, neuroendocrine/immune balance, metabolite balances, biochemical pathway speeds balances, etc [Booth et al, 2012; Gu et al. 2012; Schwalfenberg 2012; Lee et al. 2009]. If a chronic disease is caused by an imbalance, an effective treatment must be used to correct the imbalance. If a wrong treatment is misused to disturb an existing balance, the treatment causes a new imbalance. Even vitamins daily intakes can be both bad and good, depending on specific persons. Lowering omega 6/3 fatty acid ratio in patients who have a perfect ratio, using anti-virus drugs on a non-infected patient, over-detoxification of heavy metals, increasing calories intake on

obese patients, altering diet to correct a non-existing gut microbiata problem etc. will only harm the patient. Misapplication of a treatment to wrong patients is presumed to harm health properties.

Frequency data used in Table S7 is for convenience. In statistical analysis, beneficial effects μ is estimated by a computed average of all data points for the treatment. A treatment has negating effects if the treatment makes the measured health property of some treated patients worse. For example, the treatment actually shortens the survival times of treated patients. Negating effects may be non-obvious and do not have to carry negative signs. Negating effects bring down the statistical mean to lower the treatment effect, and thus have an effect of nullifying some or all beneficial effects for the treatment group. In our hypothetical model in Table S7, the treatment effect comes in three categories due to the unique nature of chronic diseases. However, they can have more categories.

In all well known statistical models, treatment effects are assumed to be constants because statistical analysis treats treatment's mean as a key comparative parameter. A common assumption used in the statistical model is that a same amount of treatment effect, μ, can be found on all patients in the treatment group, but not in the control group. However, this basic assumption does not hold. Statistical models assume that differences among individual data points within the treatment group are caused by uncontrollable random errors. Acceptable random errors are those that arise from drawing processes.

The averaging effects caused by indiscriminate use of treatments are unique. Differences among individual data points are not caused by uncontrollable random errors, but actually caused by complex, controllable disease mechanisms. Any of the treatments A to G in Table S7 work on different causes with distinctive response rates. Which patients will produce beneficial and adverse responses are determined by their matches. Both types of responses and amounts of responses depend on the patient health conditions and diseases causes. Detected values are not random variables. The measured health properties hop up and down along an imagined mean of the control according to disease mechanisms.

Leaving other problems aside, statistical models are not sophisticated enough to take account three kinds of responses: beneficial responses, non-responses, and negating responses. What the statistical analysis actually does is lowering treatment effect by averaging three types of responses, but treat their individual variations among data points as the experimental errors. The three known effects are completely different from the statistical model assumptions that all observations within a treatment or sub treatment are similar and their differences between individual data points are caused

by uncontrollable errors. Thus, the reality of human diseases is completely different from the statistical model, and the conclusions must be wrong except by accident.

The departure of the clinical trial from statistical models can been seen in numerous aspects. The beneficial effects and the negating effects of the treatment generally happen on different patients in the trial, and cannot be averaged. However, beneficial effects on patient A and negating effects on patient B is averaged in statistical operations. For a fungible thing, getting 5 dollars and losing 5 dollars is equivalent to getting nothing as far as an economic effect is concerned. In reality, the benefits of the treatments are not zero even though statistical mean is zero. The treatment can be used only to right patients to deliver beneficial effects, but not used on wrong patients to avoid adverse or negating effects. In reality, treatment A could deliver a 20% response rate rather than 4% if it is not indiscriminately used in a randomized trial.

Similar problems can be seen in other aspects. A dosage deduced from a 10 years old and a dosage deduced from a 70 years old can be summed up and averaged to become a statistical mean. A dose based this mean will be useful to neither the 10 years old, nor the 70 years old, and nor an imagined 40 years old. Similarly, an averaged heath property of two patients, one with a liver disease and the other with a kidney disease, can represent neither of them, and nor a patient with half a liver disease, and half a kidney disease. Similarly, the averaged health properties of many different types of cancer cannot represent any type of cancer in the world. The finding of 7.4% complete response rate of chemotherapy for later stage cancers cannot be used to predict a specific type of cancer or a specific person, but useful as a yardstick of the overall performance of medicine.

The averaging effect of beneficial and adverse responses in clinical trials cannot be eliminated by increasing the number of patients in the trial, but can be reduced or substantially eliminated in an optimization trial.

Column 6 in the above table shows that if treatments A to G are for addressing different causes in an optimization trial, their performance will be much higher as shown in column 7. Each treatment is used on a sub group of patients whose disease causes match the treatment. Since patients are not indiscriminately exposed to treatments, negating effects can be avoided. Moreover, multiple treatments may be used to treat patients with multiple causes. Thus, a patient may respond to a right combination even though the patient would not respond to any single treatment. Thus, treatment A in combination with other treatments could benefit more than 20% patients. If each treatment is tailored to specific patients, the treatment will not produce negating effects as in a clinical trial, and all treatments

could be available for use. In personalized medicine, the total benefits of all treatments are expected to be higher than assumed 55% and even minority patients will have treatments.

Due to huge differences among personal health properties and high accuracy required to characterize chronic diseases, any treatment protocol developed from a population trial is not relevant to a specific patient. Any statistical means from a large population cannot be applied to specific patients, and a statistical mean from a sufficiently similar patients cannot be used to other patients if they are not "sufficiently similar" to the sample patients. Health properties are not fungible things that can be exchanged between patients as abstract mathematical numbers. A person's health properties cannot be changed to match the means of the population or the values of another person. Since diseases are caused by different causes, the notion of using a single treatment indiscriminately on a population is clearly incorrect. The use of statistical analysis in optimization trials may be justified only if human subjects are sufficiently similar so that fluctuations in measured health properties are caused by uncontrollable factors and averaging is made to merely to get rid of such fluctuations. The use of statistical means in such cases is justified on the ground of reasonable approximation but not theoretical correctness.

The different responses of patients to different treatments have huge impacts on trial outcomes. The beneficial effects, μ, of a treatment is a value when the treatment is correctly used on right patients. When the treatment is indiscriminately used to a large population, the "statistically detected" treatment benefits are $\mu_s=\mu_b-\mu_n$, and can be expressed as $\mu_s=g\mu_b$, where g is coefficient to describe degrading effects which is caused by averaging the negating effects of the treatment. g is smaller than 1. If negating effects μ_n is equal to or even larger than true beneficial effects μ_b, μ_s is zero or negative and g is also zero or negative. Some g values can be seen by comparing values at column 5 to values at column 2. The table shows that negating effects in a randomized trial degrades the treatment mean. When the same treatment is used in an optimization trial, its treatment benefits are raised by (1/g), where g would be from nearly zero to the theoretical maximum of 1. This analysis shows that a randomized trial can massively degrade the treatment mean, and this degrading effect is especially large for treatments intended for rare disease causes.

This same analysis can be used to treat other treatment effects such as survival times or continuous health properties. The method can be used to analyze discrete health properties containing more than three categories. The g value can be estimated by using an empirical method. First, a statistically mean μ_s is determined by running a randomized trial and then determined by conducting statistical method. Then, true beneficial effects μ_b

for cause-matched patients are determined by running an optimization trial. Since the treatment is used only on patients, μ_b must be equal to larger than μ_s ($\mu_b \geq \mu_s$). Naturally, $g=\mu_s/\mu_b$ which will be in the range from 0 to 1. One should expect that statistical mean μ_s can be negative if the treatment has large negating effects. Thus, clinical trials result in rejecting the treatment even if the treatment could be a best cure for rare diseases if it were used to specific patients.

P. Comparison Between Multiple Factors Optimization Trial and Randomized Controlled Trial (A Model Study)

We disclose a multiple factors optimization model that is superior to the traditional clinical trial model.

1. Basic Model Assumptions and Two Models

In this model, a chronic disease is caused by a plurality of interfering factors, s_1, s_2,..., s_k, which can affect a health property which is used to measure the disease. It is assumed that a plurality of factors contribute to the diseases. They may be referred to as cause factors, interfering factors, or weak factors. Their effects are additive and all interaction terms among them are ignored for convenience. The measured health property may be a hazard rate, survival time, a vital life sign, a laboratory analysis parameter, etc. The effects of all factors are realized by a reasonable time internal. There is an uncontrollable error in measuring health property ε.

$\varepsilon \sim N(0, \sigma_E^2)$ (the true error)

$s_1 \sim N(\mu_1, \sigma_1^2)$ (a first interfering factor, which may be a treatment)

$s_2 \sim N(\mu_2, \sigma_2^2)$ (a second interfering factor)

....

$s_k \sim N(\mu_k, \sigma_k^2)$ (the kth interfering factor)

We will evaluate the performance of two treatment models: the classic clinical trial and a multiple factor optimization trial. For convenience, we assume that all k interfering factors have an equal effect: $\mu_1=\mu_2=\mu_k=\mu$, and variances $\sigma_E^2 = \sigma_1^2 = \sigma_2^2 = \sigma_k^2 = \sigma^2$. The variances of true error are sufficiently small relative to any of the interfering factor so that it can be ignored for convenience. Now, two different types of trials are used to evaluate a treatment for the disease.

A classical clinical trial (Model A). In a classic clinical trial, only

one treatment factor s_1 is selected as the treatment. The clinical trial is designed by randomizing human subjects so that s_2, s_3,…,s_k will be go into the error term. The treatment effect is $s \sim N(\mu, \sigma^2)$, while the apparent error term is $a\varepsilon \sim N(k\mu, (k-1)\sigma^2+\sigma_E^2)$ because the means and variances of k-1 causal and interfering factors are added into the error term. For convenience, we $\sigma^2=\sigma_E^2$ so that σ_E^2 can be eliminated to arrive $a\varepsilon \sim N((k-1)\mu, k\sigma^2)$. In this case, the treatment effect is only a fraction of the apparent error kμ. It is anticipated that the treatment effect will not be "found" due to the large mean and large variances of the apparent error term.

An optimization trial (Model B). All known factors s_1, s_2,…, s_k are used to optimize the effect of treatment s_1. A mini trial is designed with all k factors controlled for human subjects. All factors, s_2, s_3,…,s_k, are used in the treatment group, but not used in the control group. In an exemplar trial for studying a cancer treatment, high omega 6/3 ratios in the treatment group are corrected, but not in the control group; lack of dietary fiber intakes are corrected in the treatment group, but not in the control group, and toxic metals are detoxified in the treatment group, but not in the control…. All the k factors are controlled in the trial. When all relevant health properties are well controlled, patients are "sufficiently similar" so that summing up and averaging health properties does not amount to "averaging two different things."

In such an optimization trial, the total treatment effect, $s_t \sim N(k\mu, k\sigma^2)$, is k times larger. The apparent error term is much smaller because all k-1 interfering factors are separated and thus dropped out from the error term. The apparent error is $a\varepsilon \sim N(0, \sigma_E^2)$. All interfering factors work in a similarly adverse way within the control group, and work in a similarly beneficial way within the treatment group.

Assuming that the true error σ_E^2 is close to σ^2, the apparent error can be expressed as $\varepsilon \sim N(0, \sigma^2)$. Compared with the classical trial, treatment effect of the optimization trial is increased by k times while variances for the apparent error are decreased from $k\sigma^2$ to σ^2. It is anticipated that the trial has much higher sensitivity to "find" the total effect of all casual and interfering factor. k is the number of interfering factors plus the treatment s.

2. Performance Differences Between a Randomized Trial and An Optimization Trial

Now, we estimate how the optimization trial will improve the performance of hypothesis test results in three situations. In the analysis below, the negating effects caused by averaging is ignored for convenience.

(1) In conducting a hypothesis test using two means (see Section F). z statistic is computed by using the following equation:

$$Z_{0.05}=(\bar{X}-\bar{Y})\div\sqrt{\left(\frac{S_t^2}{n1}+\frac{S_c^2}{n2}\right)}$$

By using the optimization trial, X-Y is increased by k times. s_t^2 is the variances within the treatment group and s_c^2 is the variances of the control group. The variances of the error term are reduced from $k\sigma^2$ to σ^2. This means that the Z statistic is increased by k*√k. A similar result can be found for T statistic when the sample sizes are small. T statistic is computed by using following equation:

$$t_{0.05}(\text{n-2})=(\bar{X}-\bar{Y})\div\left(Sw*\sqrt{\frac{1}{n1}+\frac{1}{n2}}\right)$$

By using the optimization trial, X-Y is k times larger, while s_w is reduced by about √k times, the total gain is estimated to be k*√k. If optimization is done with one treatment factor plus 9 interfering factors (k=10), the total increase in T statistic is about 32 times.

(2) In conducting a hypothesis test using paired data (see Section G), using the following equation:

$$t_a=|\bar{X}|\frac{\sqrt{n}}{s}$$

The |X| is expected to be k times larger, while the standard error, s, is reduced by about k times, the total increase in T statistic is estimated to be k*√k.

(3) In conducting F-test, the F statistic is determined by following

$$F_\alpha(s-1,n-s)=\frac{(n-s)S_1}{(n-1)S_E}$$

In the above equation, s denotes levels which may be viewed as yes and no two levels. Assuming that the effect of s_1 is same as the error term and each of the interfering factors, the treatment effect is increased by k times while the error term s_E is decreased by √k times. So, the F statistic is also raised by about k*√k.

A treatment for a chronic disease may contain one, tens or more factors. In cancer cases, this is not an unreasonable number. However, not every known factor is relevant to any particular cancer patient. A plurality of factors may affect cancer outcomes but the specific mixing of specific causal or interfering factors are unique for each patient. For each patient, a right set of factors must be selected. Correcting “a problem” that does not exist in a patient can only have a negative effect.

It is well known that a large number of known factors affect cancer outcomes by different degrees. Exercises are found to have huge impacts (28%–44% reduced risk of cancer-specific mortality found in the review by Cormie et al. 2017. Since other factors affect both exercise group and non-exercise group, the true benefits of exercise may be underestimated. Emotional states have been found to have great impacts as they always exist among all cancer patients. Unhealthy diets affect patients in different ways and clinical trials yield only average effects. Pollutants are found to be weak because they are not always causes of cancer in all patients. Thus, true effects of diets and pollutants are brought down by averaging population data: that is average of good response, neutral response and adverse responses. It is expected that diets and pollutants have greater impacts on certain patients who happen to have such problems. It can be reasonably expected that if multiple relevant factors are selected and used in an optimization trial, a much larger treatment effects will be found.

If the effect of a factor is a constant with variances being close to zero, then variances of the error term does not depend on the number of cause factors used in optimization. In this extremely unlikely scenario, optimization with k factors can still raise T statics, Z statistic and F statistic by their additive effects, synergistic and interactive effects, and thus raises treatment benefits by great margins. However, many factors such as diets, nutrients, pollutants, exercise patterns affect diseases in a random fashion and are expected to work in different degrees. Even if human subject are randomized, those factors affect some patients beneficially, have no effect on some patients, affect others adversely; and if they do work, they may work by different degrees. If they are not controlled, they must raise the variances of the error term. Thus, an optimization trial can reduce error variances and improve ability to detect treatment effects.

Traditional clinical trials have more serious bias in studying toxins. Known toxic compounds are in the order several thousands. If one hundred of similar toxic substances are studied together, such optimization trial is able to detect harmful effects more than the current trial focusing on a single compound. Z statistic, T statistic, and F statistic would be 1000 times larger than that for studying a single substance. This implies that population study is an improper approach for assessing toxic compounds. It is very possible that each single compound will escape from being caught, but any of many combinations of the compounds would cause detectable damages to personal health.

The problem addressed in the study is well known in statistics as general principle. What is omitted is that, in clinical trials, expected variances are sufficiently large to result in consistent failure to recognize weak treatment effects. Even though the statistical mean of the treatment

will be centered at the mean, with the variances approaching zero when the number of patients in a clinical trial is sufficiently large. The three gains, (1/g)* k*√k, are respectfully for avoiding negating effects, addition of effects of multiple interfering factors, and reduced variances from controlling interfering factors, cannot be corrected by increasing human subject numbers.

√k is based on assumption that $\sigma^2=\sigma_E^2$ (we use one weak factor as true error that cannot be addressed). This may be viewed as the least improvement factor that impact test outcome by the error variance. If $\sigma^2>\sigma_E^{2,}$ the actual factor should be larger than √k. If all of causal and interfering factors have different effect and different variances, then the difference can be found by computing the actual sums of all causal and interfering factors and the variances for both types of trials. Big difference is expected.

Due to the similar statistical logic behind all hypothesis tests or confidence intervals, the same trend should be seen for hypothesis tests using other distributions. The root cause is a breach of the implied presumption that the sum of all experimental errors must be much smaller than the treatment effect.

3. Large Sample Size Does not Affect Relative Merits

Finally, we determine if the relative disparity in performance between randomized clinical trials and optimization trials can be changed by increasing sample sizes.

The basic model: a treatment is administered on a treatment group and is compared with a control group. The true experimental error is sufficiently small so that it can be neglected. Thus, errors within the treatment and within the control are mainly caused by one or more co-causal and interfering factors. Z statistic can be computed by using the following equation.

$$Z-Stastic=(\bar{X}-\bar{Y})\div\sqrt{\left(\frac{\sigma_t^2}{n1}+\frac{\sigma_c^2}{n2}\right)}$$

X is the treatment's mean and Y is the control's mean. For convenience, we use equal sample sizes (n1=n2=n), take Y as a zero. The error within a treatment and the control are treated as equal so that $\sigma_t^2=\sigma_c^2=\sigma_E^2$ so that Z statistic becomes:

$$Z\,Statistic=(\bar{X}/\sigma_E)\sqrt{\frac{n}{2}}$$

If Z statistic > $Z_{0.05}$=1.645, reject the null hypothesis. This condition leads to the following equation:

$$\bar{X}/\sigma_E > Z_\alpha\sqrt{\frac{2}{n}} \quad (17)$$

Whether a statistical analysis will correctly determine the treatment effect would depend on the ratio of ($\bar{X}/\sigma_E$).

Table S8 Treatment $\bar{X}$ to Error Ratios ($\bar{X}/\sigma_E$) That Could Be Enough to Reject the Null Hypothesis at an α Level for Different Sample Sizes

Sample Size n for Each Arm	α Level (Preset value)	Z_α Value from a Table	$(\sqrt{(2/n)})*Z_\alpha$	$\bar{X}$ Must be Larger Than Below Values to Reject Ho Hypothesis
50	0.05	1.645	0.20	$0.20*\sigma_E$
100	0.05	1.645	0.14	$0.14*\sigma_E$
1000	0.05	1.645	0.045	$0.045*\sigma_E$
10000	0.05	1.645	0.014	$0.014*\sigma_E$

The table shows that the sensitivity of hypothesis tests increases with sample size n. For a treatment intended for chronic diseases, X is expected to be vary small while a large number of co-causal and interfering factors can enlarge σ_E or σ_E^2. As long as the ratio is not larger than those shown in column 5, the hypothesis test outcome will be wrong.

The averaging of beneficial effects and negating effects may cause the statistically determined treatment effect $\bar{X}$ to reach zero or negative. Under this circumstance, the outcomes of randomized clinical trails are wrong. Sample size cannot alter relative performance differences between randomized trials and optimization trials because the sensitivity gain does not depend on sample size. The first term, (1/g), is stabilized by large sample sizes; the second term k depends on how well each of the interfering factors is controlled in the optimization trial. The third term (√k) does not depend on sample size (even though, the standard error of the apparent errors is approaching zero when n is approaching infinity).

REFERENCES

Altman DG. Randomisation. BMJ 1991; 302: 1481-2.

García-Pérez MA, Statistical conclusion validity: some common threats and simple remedies. Front Psychol. 2012; 3: 325.

Hart PD. A change in scientific approach: from alternation to randomised allocation in clinical trials in the 1940s. BMJ. 1999 Aug 28;319(7209):572–573.

Holt, GB (2016). Potential Simpson's paradox in multicenter study of intraperitoneal chemotherapy for ovarian cancer. Journal of Clinical Oncology, 34(9), 1016-1016.

Fleiss JL, Levin B, Park MC. A statistical Methods for Rates and Proportion. 3rd ed. Hoboken NJ: John Wiley and Sons; 2003. How to randomize.

Franks A, Airoldi E, Slavov N. (2017). "Post-transcriptional regulation across human tissues". PLOS Computational Biology. 13 (5): e1005535. doi:10.1371/journal.pcbi.1005535. ISSN 1553-7358.

Gu H, Tang C, Yang Y. Psychological stress, immune response, and atherosclerosis. *Atherosclerosis*. 2012;223(1):69–77.

Kalish LA, Begg GB. Treatment allocation methods in clinical trials a review. Stat Med. 1985;4:129–44.

Lee SJ, Trostel A, Le P et al. Cellular stress created by intermediary metabolite imbalances. Proc Natl Acad Sci U S A. 2009 Nov 17; 106(46): 19515–19520.

Legumes C.R. lemons and streptomycin: A short history of the clinical trial. CMAJ. 2009;180:23–24.

Roussas GG. A Course in Mathematical Statistics (2nd Ed). Academic Press. 1997; 327-375;440-462.

Schwalfenberg GK. The Alkaline Diet: Is There Evidence That an Alkaline pH Diet Benefits Health? J Environ Public Health. 2012; 2012: 727630.

Schul KF, Grimes DA. Allocation concealment in randomized trials: Defending against deciphering. Lancet. 2002;359:614–8.

Twyman RA. A brief history of clinical trials. The Human Genome. 2004. Sep, http://genome.wellcome.ac.uk/doc_WTD020948.html. Accessed on Oct 2009 [this article address is changed].

MRC Streptomycin in Tuberculosis Trials Committee. Streptomycin treatment of pulmonary tuberculosis. BMJ. 1948;2:769–83.

Wang B, Ogburn EL, Rosenblum M. Analysis of covariance in randomized trials: More precision and valid confidence intervals, without

model assumptions. Biometrics. 2019 Apr 22. doi: 10.1111/biom.13062.

Wagner CH. "Simpson's Paradox in Real Life". The American Statistician. February 1982.36 (1): 46–48.

Wikipedia (1). List of incurable diseases. https://en.wikipedia.org/wiki/List_of_incurable_diseases. Last assessed on July 2, 2019.

Wikipedia (2). Sum of normally distributed random variables https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables. Last assessed on July 2, 2019.