Medical Product Safety: Biological Models and Statistical Methods

Tze Leung Lai

Department of Statistics, Stanford University

June 18, 2017

Table of Contents

Benefit-risk

BR evaluation involves multiple sources of evidence and different disciplines, from clinical medicine to statistics, and policy. Available clinical trial data are usually quite limited and need to be supplemented by other sources of epidemiological data and the evaluations need to continue into the post-marketing setting.

Benefit-risk

BR evaluation involves multiple sources of evidence and different disciplines, from clinical medicine to statistics, and policy. Available clinical trial data are usually quite limited and need to be supplemented by other sources of epidemiological data and the evaluations need to continue into the post-marketing setting.

- Examples
- Ingredients for BR evaluation
- Multi–criteria statistical decision theory
- BR in clinical trial data

Benefit-risk: examples

- Tysabri was the first monocolonal antibody that was approved by the FDA in 2004 for the treatment of multiple sclerosis, but was removed from the market after two patients taking the drug developed progressive multifocal leukoencephalopathy (PML). After withdrawal from the market, an FDA advisory committee discussed and concluded that benefit is greater than risk. Based on the BR evaluation, the drug was back with limited use.
- Lorcaserin is a selective serotonin receptor agonist that regulates appetite and reduce food intake. It was shown not to reach the desired effect in short term, but long-term BR assessment shows its benefits outweighs its risk, so FDA eventually voted 18 to 4 to approve the drug.

Benefit-risk: ingredients

- Planning process:
  BRAT framework help B-R decisions; FDA also published guidance; EMA has adopted the PrOACT-URL decision-making framework.
- Qualitative and quantitative evaluations:
  Subjectivity is unavoidable in both quanlitative and quantitative evaluations.
- Benefit-risk formulations:

$$S = \sum_{k=1}^{K} w_k (P_{0k} - P_{1k}),$$

  K: number of outcomes; P_{0k} (P_{1k}: proportions of patients in control (treatment) responding; w_k: relative importance of the outcomes.
- A multidisciplinary approach incorporating multiple perspectives

Benefit-risk: Multi-criteria statistical decision theory

- MCDA (multi-criteria decision analysis):

$$S = \sum_{k=1}^{K} w_k (x_k - v_k^{(L)}) / (v_k^{(H)} - v_k^{(L)}),$$

$v_k^{(L)}$ ($v_k^{(H)}$): low(L) and high(H) weights for criterion k; x_k: the criterion value; w_k: weight of k.

- SMAA (stochastic multi-criteria acceptability analysis): two types of uncertanty: (a) sampling variability in estimated effect sizes, (b) the weights to scale relative importance of different criteria.

$$S_j(x^j, w) = \sum_{k=1}^{K} w_k u(x_k^j),$$

where $u(x_k^j) = (x_k^j - v_k^{(L)}) / (v_k^{(H)} - v_k^{(L)})$. This formulation incorporates uncertainty by assuming the criteria and the weights to be random variables

Benefit-risk: Multi-criteria statistical decision theory

- SMDM (stochastic multi-criteria discriminatory method) For two treatments A and B, $\Delta(w) = S_A(x^A, w) - S_B(x^B, w)$ is the difference in scores for the values of the K-dimensional vectors x^A and x^B and weight the vector w. Favor A if $\Delta(w) \geq \text{cutoff}(w)$ and to favor B otherwise, with cutoff(w) given by the following "discrimatory probabilities":

$$d_A = P\{\Delta \geq \text{cutoff}(w)\}, d_B = P\{-\Delta \geq \text{cutoff}(w)\}.$$

Benefit-risk: clinical trial data

- Chuang-Stein (1994) derives a response score for each patient by reducing the observed benefit response by observed risk score. The safety endpoints are divided into J classes and L_j levels of severity for each class. Intensity weights w_{jk} can be assigned to each level for each class of safety endpoints, $j = 1, \ldots, J$ and $k = 1, \ldots, L_j$. A safety (or risk) score for patient i using $r_i = \sum_{j=1}^{J} \sum_{k=1}^{L_j} w_{jk} I_{ijk}$, where $I_{ijk} = 1$ is patient i has experienced side effect j at intensity level k, and $I_{ijk} = 0$ otherwise. The efficacy score for patient i is discounted by r_i for the risk-adjusted benefit score

$$e_i^* = e_i - f r_i,$$

where f is a proportionality constant assigning penalty to the side effects profile and depends on the severity of the underlying disease being treated.

Design and analysis of clinical trials

In early-phase clinical trials, the evaluation of safety is mostly exploratory with a focus on serious adverse reactions to the product. These early first-in-human trials are conducted to identify a dose range and to gain preliminary data on safety and the pharmacokinetic properties of the candidate drug. In later phases of clinical development programs, the safety profile is characterized more fully using larger numbers of patients. Some clinical trials may be designed with specific safety hypotheses.

Design and analysis of clinical trials

In early-phase clinical trials, the evaluation of safety is mostly exploratory with a focus on serious adverse reactions to the product. These early first-in-human trials are conducted to identify a dose range and to gain preliminary data on safety and the pharmacokinetic properties of the candidate drug. In later phases of clinical development programs, the safety profile is characterized more fully using larger numbers of patients. Some clinical trials may be designed with specific safety hypotheses.

- Dose-escalation in phase I
- Safety considerations in phase II, III and phase II/III clinical trials
- Clinical trial designs with both efficacy and safety endpoints
- Analysis of safety data from clinical trials

Dose-escalation in phase I

Determine DLT (dose-limited toxicity) and MTD (maximum tolerated dose) through some rules and models:

- Rule-based design: not rely on pre-specified parametric or non-parametric models or curves, to escalate or de-escalate dose with a fraction of preceding dose, depending on the presence or absence of dose limiting toxicity among the previous cohorts of treated subjects. e.g. standard traditional 3+3 design and the accelerated titration design.
- Model-based designs: apply pre-defined statistical models or curves to toxicity data to estimate the probability of dose-limiting toxicity. Can be formulated using Bayesian framework in which the posterior probability of toxic response is updated with data from patients enrolled in each dose level. (O'Quigley et al. (1990))

Dose-escalation: model-based designs

Examples of model-based designs:

- CRM (continual reassessment method): the toxicity response is modeled through a dose-toxicity function that is continually updated using data collected and hence is in turn used to determine for the next subject the dose level that is close to the target toxicity probability threshold.
- EWOC (escalation with overdose control): modify CRM that may cause unnecessary exposure of subjects to high toxic doses by imposing additional measures to prevent future subjects from being exposed to high toxic doses. EWOC method assesses the probability of exceeding the MTD for each higher dose after each patient, prohibiting dose escalation if this probability exceeds some pre-specified value.
- Bartroff and Lai (2010, 2011) proposes to use Bayesian sequential designs; also see Whitehead and Brunier (1995), Haines, Perevozskaya and Rosenberger (2003) .

Safety for the design of phase II and III studies

- Challenges in phase II and phase III trials: (a) demonstrating serious safety concerns with statistical power needs much larger sample size; (b) lack of evidentiary standards for evaluating safety; (c) prior evidence on safety on an agent not strong enough to prompt phase II and III trials designed to address the serious safety issues.
- Conditioning on rare adverse events: randomized two-arm trial with $1 : r$ randomization ratio; λ_0 (λ_1): incidence rates; n_0 (n_1): observed number of events from control (treatment); constant incidence rate $=$ Poisson arrivals of AEs. n_1 follows a binomial distribution $Bin(n, \pi)$, where $\pi = r\lambda_1/(\lambda_0 + r\lambda_1)$ .The relationship between the relative risk $R = \lambda_1/\lambda_0$ and the binomial probability π is

$$R = \frac{\pi}{r(1-\pi)} \text{ and } \pi = \frac{rR}{1+rR}.$$

Sequential conditioning methods and efficient GLR tests

- Use repeated significance test that terminates after n intussusception cases and declares the vaccine to be unsafe if $P\{Bin(n, p_0) \geq \#_n(V)\} \leq 0.025$, wehre $\#_n(V)$ denotes the number of vaccine cases among the n cases. Declare the vaccine to be safe if $P\{Bin(n, p_0) \leq \#_n(V)\} \leq 0.025$, where $p_1 = 11/10 p_0$ corresponds to a 10-fold increase in risk for the vaccine group.
- MC simulations show that the probability for the study to stop with "no increase risk using vaccine" is 0.94 for a vaccine with no increased risk, and the probability for the study to stop with a "increased-risk vaccine" conclusion is almost 1 for relative risks of 6 or greater. This "conservativeness" is good for safety evaluation.

Sequential tests without conditioning on n

- Next, an innovation without conditioning on the total number of events, which makes conventional sequential tests applicable. The limit of binomial distribution is Poisson.
- Arrivals of adverse events follow a Poisson process, with rate λ_V for vaccine (V) and λ_C for placebo (C). Test $H_0 : \lambda_V/\lambda_C \leq 1$ versus $H_1 : \lambda_V/\lambda_C \geq \gamma$, where $\gamma > 1$. Let $p = \frac{\lambda_V}{\lambda_V+\lambda_C}$. Shih et al (2010) propose to use a sequential generalized likelihood ratio (GLR) test

$$\tau = \inf\{n \geq 1 : l_{n,0} \geq b \text{ or } l_{n,1} \leq a\},$$

which is asymptotically efficient for testing $H_0 : p \leq p_0$ versus $H_1 : p \geq p_1$.

Sequential conditioning methods and BLN(2004)

- Bartroff, Lai and Narasimhan (2014) (BLN) propose an integrated approach: a joint efficacy-toxicity model is chosen to model toxicity y_i and efficacy z_i , and a phase I design is chosen to estimate MTD. In phase II, a group-sequential GLR test of $H_0 : P(z = 1|x = \gamma) \leq p_0$, rather than $K_0 : p(\widehat{\gamma}_0) \leq p_0$ is used. The MTD estimate $\widehat{\gamma}$ is updated at each stage and always dose patients at the current estimate.

Incidence rate: a better measurement

- Safety data are commonly summarized using the **crude incidence rate**: $\frac{\#\text{subjects with adverse events}}{\#\text{subjects in each treatment group}}$. Its validity requires strong assumptions of randomness of patient discontinuation and of constant hazard rates within groups over time.
- A remedy: **Exposure-adjusted incidence rate** (EAIR): calculate the exposure time period of each patient and then divide the number of subjects with AEs by the total exposure time for all patients.

Various confidence intervals

- CI based on Wald's approximation and moment
  - Conventional Wald's CI: $\hat{\lambda}_1 - \hat{\lambda}_2 \pm Z_{1-\alpha/2}\hat{\sigma}$.
  - CI with moment incorporated in:
    $\hat{\lambda}_1 - \hat{\lambda}_2 + \delta \pm Z_{1-\alpha/2}\sqrt{\hat{\sigma} + \delta^2}$, where
    $\delta = Z^2_{1-\alpha/2}(1/T_1 - 1/T_2)/2$.
- CI based on variance estimate recovery:
  Li et al. (2014) constructed the confidence intervals on the difference of two Poisson rates based on the recovered variance estimates of λ_1 and λ_2. Let L and U be the lower and upper bounds of the confidence intervals based on Wald's approximation. Then, L and U can be viewed as the minimum and maximum values of θ satisfying

$$Z^2_{1-\alpha/2} = \frac{(\hat{\lambda}_1 - \hat{\lambda}_2 - L)^2}{\hat{\sigma}^2} \text{ and } Z^2_{1-\alpha/2} = \frac{(U - \hat{\lambda}_1 + \hat{\lambda}_2 - L)^2}{\hat{\sigma}^2}.$$

Various confidence intervals

Let (l_1, u_1) and (l_2, u_2) denote the two-sided $100(1-\alpha)\%$ confidence intervals on λ_1 and λ_2, respectively. Then the variance estimates of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ can be recovered. The confidence intervals on θ is given by

$$\begin{cases} L = \hat{\lambda}_1 - \hat{\lambda}_2 - \sqrt{(\hat{\lambda}_1 - l_1)^2 + (u_2 - \hat{\lambda}_2)^2} \\ U = \hat{\lambda}_1 - \hat{\lambda}_2 - \sqrt{(u_1 - \hat{\lambda}_1)^2 + (\hat{\lambda}_2 - l_2)^2}. \end{cases}$$

Various confidence intervals

- Other CI types:
  - CI based on parameter constraint
  - CI with stratification
  - ...

Limitations and other adjustments of methods

- Limitations: large number of scales, inadequate statistical power, non-meaningful p-values, heterogeneity in adverse event reporting...

Limitations and other adjustments of methods

- Limitations: large number of scales, inadequate statistical power, non-meaningful p-values, heterogeneity in adverse event reporting...

Some other adjustments of methods:

- Integrated Summary of Safety (ISS): a compound has multiple disease indications for countries and populations with diversified background
- Development Safety Update Report (DSUR): to present annual review from all clinical trias
- Crude and exposure-adjusted incidence rates

Multiplicity in the evaluation of clinical safety data

Due to new unanticipated effects, there is potential for drawing false positive conclusions and the need for understanding the multiplicity aspects in safety signal detection. Safety assessment continues into the post-marketing phase initially with clinical trials designed specifically to address possible safety issues. We describe both frequentist error-controlling methods and Bayes (in particular, empirical Bayes) methods that have been developed to address multiplicity in the evaluation of clinical safety data.

Multiplicity in the evaluation of clinical safety data

Due to new unanticipated effects, there is potential for drawing false positive conclusions and the need for understanding the multiplicity aspects in safety signal detection. Safety assessment continues into the post-marketing phase initially with clinical trials designed specifically to address possible safety issues. We describe both frequentist error-controlling methods and Bayes (in particular, empirical Bayes) methods that have been developed to address multiplicity in the evaluation of clinical safety data.

- Illustrative example
- Multiplicity and FDR
- Double FDR and FDR with discrete statistics
- Berry and Berry's hierarchical mixture Bayes model
- Gould's Bayesian screening

Multiplicity: example

- Data: MMRV vaccine trial, treatment receives a measles, mumps, rubella, varicella (MMRV) combination vaccine, control receives MMR on Day 0 and optional V 42 days later. $N_T = 148, N_C = 132$.
- MedDRA and body systems: The adverse events are coded using a standard dictionary (e.g., MedDRA) and classified into groupings by body systems. The MMRV dataset consists of 40 adverse event types which are categorized into 8 body systems
- Category of Tier 1,2,3 events; we focus on Tier 2 events.

Multiplicity: FDR

- FWER (family-wise error rate) control ($P(V \geq 1) \leq \alpha$) versus FDR (false discovery rate) control ($E(V/R) \leq \alpha$); V: number of true null hypotheses rejected, R: number of all hypotheses rejected
- FDR-controlled procedure has higher power than FWER-controlled procedure
- FWER control leads to Bonferroni procedure; FDR control leads to BH (Benjamini & Hochberg) procedure.

Multiplicity: FDR and DFDR

- BH procedure: FDR adjusted P-value

$$P_{[m]} = P_{(m)}, P_{[j]} = \min\{P_{[j+1]}, m/j \cdot P_{(j)}\},$$

  for $j \leq m - 1$.
- Mehrotra and Heyse (2004) propose double FDR (DFDR) control which incorporates structure of body systems. Need clever resampling method.
  - DFDR: first apply BH procedure on body system level with level α_1, and then apply BH procedure on individual hypotheses in selected body systems with level α_2.
  - Can fix $\alpha_1 = \alpha_2/2$, but does not control FDR. Use resampling technique can give better FDR control at level α_2.

Multiplicity: FDR and DFDR

- Mehrotra and Adewale (2012) propose an adjustment to DFDR procedure developed by Mehrotra and Heyse (2004):
  - First, apply BH procedure on the **adjusted smallest P-value** for each body system level with level α_1. Then **put all individual hypotheses from selected body systems together** and apply BH procedure **once** with level α_2.
  - Mehrotra and Adewale (2012) fix α_1, α_2 without using any resampling methods. This does not control FDR at level α_2 theoretically, even under independence assumption.

Multiplicity: comparisons of 2004 & 2012 DFDR

A detailed comparison...

- Original (2004) DFDR procedure:
  - Step 1: Let $p_i = \min(p_{i1}, p_{i2}, ..., p_{ik_i})$ denote the representative p-value for body system i, and let $\tilde{p}_i$ denote the corresponding BH FDR-adjusted p_i.
  - Step 2: Let $\tilde{p}_{ij}^{(i)}$ denote the FDR-adjusted p_{ij} obtained by applying a BH FDR adjustment to the k_i p-values *within* body system i
  - Reject AE_{ij} if $\tilde{p}_i \leq \alpha_1$ and $\tilde{p}_{ij}^{(i)} \leq \alpha_2$.
- (2012) new DFDR procedure:
  - Step 1: Apply BH FDR adjustment to the $p_i^* (1 \leq i \leq s)$ values, where $p_i^* = \min(\tilde{p}_{ij}^{(i)}; 1 \leq j \leq k_i)$, and let $\tilde{p}_i^*$ denote the FDR-adjusted p_i^*.
  - Step 2: Let $F \equiv \{p_{ij} | \tilde{p}_i^* \leq \alpha\}$. Apply a single BH FDR adjustment to the p-values in F, and let $\tilde{p}_{ij}^{(F)} =$ FDR-adjusted $p_{ij} | p_{ij} \in F$.
  - Reject AE_{ij} if $\tilde{p}_i^* \leq \alpha$ and $\tilde{p}_{ij}^{(F)} \leq \alpha$.

Multipilicity: New appraoch

- A combination of the new DFDR procedure (2012) and careful resampling might be a better choice
- Lai, Miao, Tsang (2017): alternative approach to "divide and conquer" idea underlying DFDR for multiple hypothesis testing and post-selection inference.
  - Insights:...

Multiplicity: FDR control for discrete data

- Heyse (2011) improves power even more by adjusting FDR for discrete test statistics:
  - Let $Q_i(p)$ denotes the largest achievable P-value $\leq p$ for hypothesis $i = 1, ..., m$. $Q_i(p) = 0$ if no such P-value exists. The adjusted P-values for discrete data is:

$$P_{[m]} = P_{(m)}, P_{[j]} = \min\{P_{[j+1]}, \sum_{i=1}^{m} Q_{(i)}(P_{(j)}/j\},$$

    for $j \leq m - 1$.
  - The discrete-adjusted P-value ($Q_{(i)}(P_{(j)}) \leq P_{(j)}$) is smaller than or equal to the originally BH-adjusted P-value, ($Q_{(i)}(P_{(j)}) = P_{(j)}$).

Multiplicity: FDR and DFDR

Table: Smallest B&H Adjusted P-value From Each of the 8 Body Systems

BS	# AEs	Representative AE Description	Group 1 $N_1 = 148$	Group 2 $N_2=132$	Unadjusted P-value	Adjusted P-value
01	5	Asthenia/Fatigue	57	40	0.1673	0.6248
03	7	Diarrhea	24	10	0.0289	0.2026
05	1	Lymphadenopathy	3	2	1.0000	1.0000
06	1	Dehydration	0	2	0.2214	0.2214
08	3	Irritability	75	43	0.0025	0.0075*
09	11	Bronchitis	4	1	0.3746	0.9447
10	9	Rash	13	3	0.0209	0.1745
11	3	Conjunctivitis	0	2	0.2214	0.6641

Multiplicity: Hierarchical Bayes mixture model

- Berry and Berry's 3-level hierarchical Bayes model: θ_{bi} is the logarithm of the odds ratio of the adverse event probability for treatment (Group 2) to that for control (Group 1):

$$\theta_{bi} = \log(p_{bi,2}/(1-p_{bi,2})) - \log(p_{bi,1}/(1-p_{bi,1})),$$

$p_{bi,1}$ and $p_{bi,2}$: adverse event probabilities for Group 1 and Group 2; see Table on next slide. There is positive, albeit small, posterior probability that $\theta_{bi} < 0$ in the Bayesian model. The first level of the Bayesian hierarchical mixture model assumes that $\theta_{bi} = 0$ with probability π_b and is normally distributed with probability $1 - \pi_b$. The second and third levels of the hierarchical specification gives the prior distributions of π_b and of the mean and variance of the normally distributed component of the mixture model at the first level.

Multiplicity: Hierarchical Bayes mixture model

- Bayesian specification attempts to model "the existing structure and the available information" among types of adverse events (AEs) "explicitly depending on their body systems," thus "borrowing information across types of AEs." Hence, "this is different from conclusions of more traditional multiple comparison methods in which only the number of types of AEs under consideration matters," as in the FDR and DFDR control methods. There is only one type of AE (irritability in body system 8) with a high value (0.78) for the posterior probability of $\theta_{bi} > 0$. This AE type also has the smallest P-value (0.003) for Fisher's exact tests.
- This 3-level mixture model has "borrowing" and "shrinking" effect within each body system.
- We develop new methods to significantly ease the MCMC computation

Multiplicity: Hierarchical Bayes mixture model

Table: Fisher's 2-sided P-values (with asterisks **if** $<$ **0.1** and posterior probabilities under the 3-level hierarchical Bayesian model

			2-sided	Posterior probability	
b	i	Type of AE	P-value	$\theta_{bi} > 0$	$\theta_{bi} = 0$
...	...	...	...	...	...
8	1	Crying	0.500	0.185	0.655
8	2	Insomnia	1.000	0.153	0.661
8	3	**Irritability**	0.003*	**0.780**	0.214
9	1	Bronchitis	0.375	0.059	0.900
...	...	...	...	...	...

Multiplicity: Bayesian screening

- Gould's Bayesian screening:
  Due to the fact that "testing hypotheses about treatment group differences in adverse event incidence when the adverse events have not been identified in the study protocol amounts to using observed data to test hypotheses that are generated by the same data." He advocates a Bayesian screening approach that "provides a direct assessment of the likelihood of no material drug-event association and quantifies the strength of the observed association" for the Tier 2 AEs of the control and treatment groups.
- The model: $p_{bi,2}$ equals to $p_{bi,1}$ with probability π and has a Beta distribution that is independent of the Beta distribution for $p_{bi,1}$ with probability $1-\pi$, and that π also has a Beta distribution. The parameters of the Beta prior distributions are determined from the data so as to strike a good balance between sensitivity and specificity of the classifier.

Multiplicity: Bayesian screening

- Screening rule:
  log odds ratio $\theta_{bi} \leq \theta^*$ for classifying the observed AE as safe, and flagging concerns if $\theta_{bi} > \theta^*$.
- Model (continued): when sample size is large, can model AE types as Poisson distribution. Then compare risk ratio λ_1/λ_0.
- Posterior probabilities are much easier to compute than Berry and Berry's three-level hierarchical model.
- This is implicitly related to empirical Bayes and use local FDR, which is useful tool to quanitify and balance specificity versus sensitivity.

Causal inference from post-marketing data

Post-marketing data from clinical trials and observational studies are important for regulatory agencies “to monitor the safety of drugs after they reach the marketplace and to take corrective action if drugs risks are judged unacceptable in light of their benefits” .

- Data from phase IV clinical trials and observational data from spontaneous reporting of adverse events by users of approved drugs.
- Introduction to causal inference and associated statistical models and methods.
- Observational studies and causal inference methods in these studies.
- **Structural equation models, causal diagrams and graphical models of causal effects, using perspectives from computer science**
- Prediction, statistical learning and data-driven decisions in causal models

Post-marketing data collection

- Upon regulatory approval, a medical product is continuously investigated through post-marketing studies. Depending on the type and amount of evidence, the objective and design considerations of post-marketing studies could differ substantially.
- If a product targeting a rare disease for which no efficacious therapies are available is approved based on a limited amount of evidence on clinical efficacy and safety, then the regulatory agency may require the manufacturer to conduct a randomized controlled trial to further confirm the product's benefit and risk. But if a product with a well-established safety profile in clinical trials is approved based on convincing evidence of efficacy, then a post-marketing surveillance program consisting of non-interventional epidemiological observational studies may be required to demonstrate long-term safety.

Post-marketing data collection

Different types of post-marketing clinical trials:

- Clinical trials with safety endpoints
- Observational pharmacoepidemiologic studies using registries
- Prospective cohort observational studies
- Retrospective observational studies

Causality, potential outcomes and counterfactuals

- Conterfactuals, potential outcomes, and Rubin's causal model:

$$PC = P\{Y(0) = 1 | E = 1, Y(1) = 1, X\},$$

where $E = 1$ represents "exposure" and Y is response variable.

- The terms "cause" and "treatment" are used interchangeably in Rubin's causal model. Each unit is potentially exposable to any one of these treatments before exposure, and has received only one treatment post-exposure. Much of the literature considers the case $\mathcal{J} = \{t, c\}$ consisting of two elements t (for treatment) and c (for control). We consider here more general $\mathcal{J}$, which are finite sets that allow for different levels for the treatment t (as in dose levels of a drug or amount of smoking for cigarette smokers). The assignment variable T_i assigns to the ith unit the cause or treatment in $\mathcal{J}$ that acts on it.

Rubin's causal model

We can write observations as $\{(Y_i(T_i), \mathbf{X}_i), i = 1, ..., n\}$, in which $\mathbf{X}_i$ is the pre-exposure covariate of the ith unit. Causal inference is about comparison of the distributions of potential outcomes $Y_i(\tau), \tau \in \mathcal{T}$ under the following two assumptions:

- *Stable Unit Treatment Value Assumption (SUTVA)*. Given the observed covariates $\mathbf{X}_1, ..., \mathbf{X}_n$, the distribution of potential outcomes of one unit is independent of the potential treatment assignments for other units.
- *Ignorability*: T_i has the same conditional distribution given $\{(\mathbf{X}_i, Y_i(\tau)) : \tau \in \mathcal{T}, 1 \leq i \leq n\}$ as that given $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)$, and $P(T_i = \tau | \mathbf{X}) \geq \epsilon$ for some $\epsilon > 0$ and all $\tau \in \mathcal{T}$ and $1 \leq i \leq n$.

Causality: Frequentist, Bayesian, and missing data approaches

- Frequentists' approach:

$$\begin{aligned}
\mathrm{E}(\overline{Y}(\tau)) &= \mathrm{E}\left\{\mathrm{E}\left[n_\tau^{-1}\sum_{i=1}^{n} Y_i(\tau)\mathbf{1}_{\{T_i=\tau\}}\middle|\mathbf{X}\right]\right\} \\
&= \mathrm{E}\left\{\mathrm{E}\left[n_\tau^{-1}\mathrm{E}(Y(\tau)|\mathbf{X})\sum_{i=1}^{n}\mathbf{1}_{\{T_i=\tau\}}\middle|\mathbf{X}\right]\right\} \\
&= \mathrm{E}\{\mathrm{E}(Y(\tau)|\mathbf{X})\} = \mu(\tau).
\end{aligned}$$

Thus, the frequentist approach to causal inference uses the usual tools of consistent and asymptotically normal estimators of the means of potential outcomes, from which confidence intervals for the mean causal effects can be derived.

Causality: Frequentist, Bayesian, and missing data approaches

- Bayesian approach:
  Dawid et al. (2016) also introduce a Bayesian approach. In principle, this approach begins with a prior distribution of a multivariate parameter comprising the probabilities of the four configurations of $(Y(0), Y(1))$ conditioned on X (since $Y(0)$ and $Y(1)$ are binary outcomes) and then derive a fully determined posterior distribution for equation in previous slide. However, they point out that this is problematic because $Y(0)$ and $Y(1)$ are never simultaneously observable and therefore the parameter describing the joint distribution of $(Y(0), Y(1))$ given X is not identifiable from the data, making the Bayesian inference highly sensitive to the specific prior assumptions made. They therefore assign a joint prior distribution for the estimable probabilities $\mathrm{P}(Y = 1|E = 1, X)$ and $\mathrm{P}(Y = 1|E = 0, X)$.

Causality: Frequentist, Bayesian, and missing data approaches

- Bayesian approach (continued):
  Rubin uses full prior specification in his Bayesian approach to causal inference but emphasizes the importance of ignorable treatment assignment mechanism for the causal inference to be insensitive to the prior distribution. In particular, the concept of ignorable treatment assignment is introduced in the 1978 paper.
  Rubin also embeds Bayesian causal inference in the broader framework of Bayesian imputation methods for missing data.

Observational studies: confounding and adjustments

- Matching, subclassification, and standardization
  - Matching, or more precisely matching samples, refers to forming a sample of size n from the set of observed $(X_i, Y_i(c)), i = 1, ..., rn$ of rn units ($r \geq 1$) assigned to the control so that the the covariate values X_i's match (in some way) the $\tilde{X}_j$'s in the observations $(\tilde{X}_j, Y_j(t)), j = 1, ..., n$, from the treatment group. Rubin proves that matching reduces bias in estimating $\mathrm{E}(Y(t) - Y(c))$ when the univariate X_i do not have the same distribution as the $\tilde{X}_j$.

Observation studies: subclassification and standardization

- Subclassification is another method of adjustment for confounding. The treatment and control groups are divided into subclasses or strata on the basis of the covariate $\mathbf{X}$, so that each subclass can be regarded as having approximately the same vaues of $\mathbf{X}$. Although the method is natural for discrete $\mathbf{X}$ that takes on a relatively small number of values, it encounters major difficulties when the covariate vector $\mathbf{X}$ is continuous and has a large number of components.
- Standardization refers to reweighting the observations for confounder control.

Other methods for observational studies

- Propensity score
  - Control for confounding via estimated propensity score
- Inverse probability weighting
- Time-dependent confounding and g-estimation
- Model-based adjustments and sensitivity analysis

DAGs and causal effects

- Directed acyclic graphs and symbolic derivation of causal effects

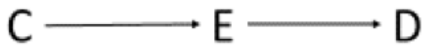

Figure: Example 1 of DAGs

A causal diagram is known as a directed acyclic graph (DAG). The graph represents three random variables (E, C, D) as *nodes* (or *vertices*). These three nodes are connected by *edges* (the arrows). C is temporally prior to E and D, and E is temporally prior to D. Two nodes are *adjacent* if there is an edge between them. A *path* between two nodes C and D is a sequence of nodes beginning with C and ending with D, in which each node is connected to the next by an edge.

DAGs

- Hernán et al(2017) characterize a causal DAG with the following criteria: (1) the lack of an arrow from one node to another can be interpreted as the absence of a direct causal effect of the two variables, relative to the other variables on the graph; (2) all common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph; and (3) any variable is a cause of its descendants.
- **Causal Markov assumption**: conditional on its direct causes, a variable V_j is independent of any variable of which it is not a cause. That is, conditional on its parents, V_j is independent of its non-descendants. This statement is mathematically equivalent to the statement that the density $f(V)$ of the variables V in DAG G satisfies the Markov factorization $f(v) = \prod_i P(X_i|pa_i)$.

Conditional independence in graphical models

- Reference: Hernán et al. (2017), Pearl (1988, 1995, 2016), Whittaker (1990), etc
- Pearl summarizes the graphical methods that are used to identify the conditional independence relationships are based on the recursive product decomposition where

$$P(X_1, ..., X_n) = \prod_i P(X_i|pa_i),$$

where pa_i stands for the realization of some subset of the variables that precede X_i in the order $(X_1, X_2, ..., X_n)$. If we construct a directed acyclic graph (DAG) in which the variables corresponding to pa_i are represented as the parents (or adjacent predecessors or direct influences) of X_i, then we have a way to determine independence...

d-separation test

- The independencies implied by the decomposition can be read off the graph using the *d-separation test* , defined as follows:
  - A path p in a DAG G is blocked by a set of nodes Z if and only if
    (a) p contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that the middle node B is in Z (that is, B is conditioned on), or
    (b) p contains a collider $A \to B \leftarrow C$ such that the collision node B is not in Z, and no descendants of B is in Z.
    If Z blocks every path between two nodes X and Y, then X and Y are d-separated conditional on Z, and thus are independent conditional on Z.

d-separation test

Geiger et al. (1990) show that there is a one-to-one correspondence between the set of conditional independence between X and Y given Z implied by the recursive decomposition, and the set (X, Y, Z) that satisfy the d-separation criteria in G.

Examples of d-separation test

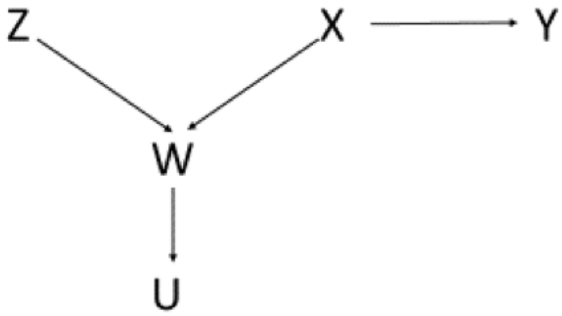

Figure: Example 1 of using d-separation to determine conditional independence

Focus on the relationship between Z and Y. If we use an empty conditioning set, they are d-separated because there is no unblocked path between them.
If we condition on W, then Z and Y are d-connected because the only path between Z and Y contains a fork (X) that is not in that set, and the only collider (W) on the path is in the set, the path is not blocked.
Z and Y are d-connected if we condition on U

Examples of d-separation test

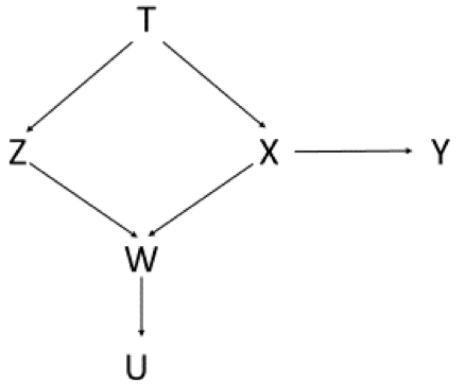

Figure: Example 2 of using d-separation to determine conditional independence

If we add another path between Z and Y, then Z and Y are conditionally dependent, because there is a path between them ($Z \leftarrow T \rightarrow X$) that contains no colliders.
If condition on T, then that path is blocked, and Z and Y become independent again.
Conditioning on $\{T, W\}$ will make Z and Y d-connected because conditioning on T blocks the path $Z \leftarrow T \rightarrow Y$, but conditioning on W unblocks the path $Z \rightarrow W \leftarrow X \rightarrow Y$.

Conclusion

- There are other issues to be solved, e.g. confounding covariates that cause the adverse events and adjustments have to be made for causality analysis. The causaility model we use is important.

Conclusions

- Techniques that can potentially be integrated to address the challenges of using safety databases are
  - Statistics methods: propensity scores, graphical models, instrumental variables, and inverse probability weighting .
  - Phamacoepidemiology: assessment of medication adherence and medication errors (or of device misuse or malfunctioning leading to device-related adverse experiences for medical devices), reporting ratios and disproportionality analysis, case-control approach and self-controlled case series.