Overview of Epidemiological Study Designs

Madhukar Pai, MD, PhD
McGill University
madhukar.pai@mcgill.ca
A major goal of epi research is causality

Epi studies measure 3 things: exposures, confounders & outcomes

Once quantified, the association between exposure and outcome is the central focus

There are many ways of evaluating the association between an exposure and an outcome: these are the different study designs
The best epi study will be one that captures the causal effect of interest with minimal distortion (error)

Adapted from: Maclure, M, Schneeweis S. Epidemiology 2001;12:114-122.
Classification of study designs (Version 8)
(Qualitative studies are not included in this scheme; categories shown are not necessarily mutually exclusive; hybrid and mixed designs are possible)

Descriptive studies
- designed to describe occurrence of disease by time, place and person

Experimential (intervention studies)
- investigator intentionally alters one or more factors to study the effects of so doing

Quasi-experimental
- investigator lacks full control over the intervention but conducts the study as if it were an experiment

Non-experimental (observational studies)
- does not involve intervention; investigator observes without intervention other than to record, count, and analyze results

Uncontrolled trials
- experimental trials without control or comparison groups (e.g. phase I/II clinical trials)

Controlled trials
- trials with control groups (e.g. phase III clinical trials)
- controlled trials can be clinical trials (unit of randomization is an individual) or community/trial trials (unit of randomization is a community or cluster)

Randomized (RCTs)
- interventions allocated randomly (all participants or clusters have the same chance of being allocated to each of the study groups)

Quasi-randomized
- allocation done using schemes such as: according to date of birth (odd or even), number of the hospital record, date at which they are invited to participate in the study (odd or even), or alternatively into the different study groups

Non-randomized
- allocation to different groups done arbitrarily (without any underlying random process)

Note: Systematic reviews and meta-analyses involve the secondary analysis and synthesis of original studies and are not considered in this classification system

Madhukar Pai, McGill University, Montreal (madhukar.pai@mcgill.ca)
Kristian Filoni, McGill University, Montreal (kristian.filoni@mail.mcgill.ca)
Experimental designs

**Uncontrolled trials**
- experimental trials without control or comparison groups (e.g. phase I/II clinical trials)

**Controlled trials**
- trials with control groups (e.g. phase III clinical trials)
- controlled trials can be clinical trials (unit of randomization is an individual) or community/field trials (unit of randomization is a community or cluster)

**Randomized (RCTs)**
- interventions allocated randomly (all participants or clusters have the same chance of being allocated to each of the study groups)

**Quasi-randomized**
- allocation done using schemes such as according to date of birth (odd or even), number of the hospital record, date at which they are invited to participate in the study (odd or even), or alternatively into the different study groups

**Non-randomized**
- allocation to different groups done arbitrarily (without any underlying random process)
Compassionate Use of Remdesivir for Patients with Severe Covid-19


BACKGROUND
Remdesivir, a nucleotide analogue prodrug that inhibits viral RNA polymerases, has shown in vitro activity against SARS-CoV-2.

METHODS
We provided remdesivir on a compassionate-use basis to patients hospitalized with Covid-19, the illness caused by infection with SARS-CoV-2. Patients were those with confirmed SARS-CoV-2 infection who had an oxygen saturation of 94% or less while they were breathing ambient air or who were receiving oxygen support. Patients received a 10-day course of remdesivir, consisting of 200 mg administered intravenously on day 1, followed by 100 mg daily for the remaining 9 days of treatment. This report is based on data from patients who received remdesivir during the period from January 25, 2020, through March 7, 2020, and have clinical data for at least 1 subsequent day.

RESULTS
Of the 61 patients who received at least one dose of remdesivir, data from 8 could not be analyzed (including 7 patients with no post-treatment data and 1 with a dosing error). Of the 53 patients whose data were analyzed, 22 were in the United States, 22 in Europe or Canada, and 9 in Japan. At baseline, 30 patients (57%) were receiving mechanical ventilation and 4 (8%) were receiving extracorporeal membrane oxygenation. During a median follow-up of 18 days, 36 patients (68%) had an improvement in oxygen-support class, including 17 of 30 patients (57%) receiving mechanical ventilation who were extrapolated. A total of 25 patients (47%) were discharged, and 7 patients (13%) died; mortality was 18% (6 of 34) among patients receiving invasive ventilation and 5% (1 of 19) among those not receiving invasive ventilation.

CONCLUSIONS
In this cohort of patients hospitalized for severe Covid-19 who were treated with compassionate-use remdesivir, clinical improvement was observed in 56 of 53 patients (68%). Measurement of efficacy will require ongoing randomized, placebo-controlled trials of remdesivir therapy. (Funded by Gilead Sciences.)

NEJM 2020
Hydroxychloroquine in patients with mainly mild to moderate coronavirus disease 2019: open label, randomised controlled trial

Weifang Tang,1,2 Zhujun Cao,3 Mingfeng Han,4 Zhengyan Wang,5 Junwen Chen,6 Wenjin Sun,7 Yaojie Wu,8 Wei Xiao,9 Shengyong Liu,10 Erzhen Chen,11 Wei Chen,12 Xiongbiao Wang,12 Jiuyong Yang,13 Jun Lin,14 Qingxia Zhao,15 Youqin Yan,16 Zhibin Xie,17 Dan Li,18 Yaofeng Yang,19 Leshan Liu,20 Jieming Qu,1,2 Guang Ning,21 Guochao Shi,1,2 Qing Xie3

DESIGN
Multicentre, open label, randomised controlled trial.

SETTING

PARTICIPANTS
150 patients admitted to hospital with laboratory confirmed covid-19 were included in the intention to treat analysis (75 patients assigned to hydroxychloroquine plus standard of care, 75 to standard of care alone).

INTERVENTIONS
Hydroxychloroquine administrated at a loading dose of 1200 mg daily for three days followed by a maintenance dose of 800 mg daily (total treatment duration: two or three weeks for patients with mild to moderate or severe disease, respectively).

MAIN OUTCOME MEASURE
Negative conversion of severe acute respiratory syndrome coronavirus 2 by 28 days, analysed according to the intention to treat principle. Adverse events were analysed in the safety population in which hydroxychloroquine recipients were
Simple, two-arm (parallel) RCT

**FIGURE 10.1**
In a randomized trial, the investigator (a) selects a sample from the population, (b) measures baseline variables, (c) randomizes the participants, (d) applies interventions (one should be a blinded placebo, if possible), (e) follows up the cohort, (f) measures outcome variables (blindly, if possible) and analyzes the results.
Cross-over RCT design

**FIGURE 11.4**
In the cross-over randomized trial, the investigator (a) selects a sample from the population, (b) measures baseline variables, (c) randomizes the participants, (d) applies interventions, (e) measures outcome variables, (f) allows washout period to reduce carryover effect, (g) applies intervention to former placebo group, (h) measures outcome variables again.
Non-experimental (observational) designs

- Cohort studies
- Case-control studies
- Cross-sectional studies
- Ecologic studies
- Diagnostic accuracy studies
Cohort study

Figure 1–15 Same cohort study as in Figure 1–13, but the ascertainment of events and losses to follow-up is done separately among those exposed and unexposed.
Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study

Fei Zhou*, Ting Yu*, Ronghui Du*, Guohui Fan*, Ying Liu*, Zhibo Liu*, Jie Xiang*, Yeming Wang, Bin Song, Xiaoying Gu, Lulu Guan, Yuan Wei, Hui Li, Xudong Wu, Jiaying Xu, Shengjiong Tu, Yi Zhang, Hua Chen, Bin Cao

Summary

Background Since December, 2019, Wuhan, China, has experienced an outbreak of coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Epidemiological and clinical characteristics of patients with COVID-19 have been reported but risk factors for mortality and a detailed clinical course of illness, including viral shedding, have not been well described.

Methods In this retrospective, multicentre cohort study, we included all adult inpatients (≥18 years old) with laboratory-confirmed COVID-19 from Jinyintan Hospital and Wuhan Pulmonary Hospital (Wuhan, China) who had been discharged or had died by Jan 31, 2020. Demographic, clinical, treatment, and laboratory data, including serial samples for viral RNA detection, were extracted from electronic medical records and compared between survivors and non-survivors. We used univariable and multivariable logistic regression methods to explore the risk factors associated with in-hospital death.

Findings 191 patients (135 from Jinyintan Hospital and 56 from Wuhan Pulmonary Hospital) were included in this study, of whom 137 were discharged and 54 died in hospital. 91 (48%) patients had a comorbidity, with hypertension being the most common (58 [30%] patients), followed by diabetes (36 [19%] patients) and coronary heart disease (15 [8%] patients). Multivariable regression showed increasing odds of in-hospital death associated with older age (odds ratio 1.10, 95% CI 1.03-1.17, per year increase; p=0.0043), higher Sequential Organ Failure Assessment (SOFA) score (5-6, 0.61-12.23; p<0.0001), and d-dimer greater than 1 μg/mL (18.42, 2.64-128.55; p=0.0033) on admission. Median duration of viral shedding was 20.0 days (IQR 17.0-24.0) in survivors, but SARS-CoV-2 was detectable until death in non-survivors. The longest observed duration of viral shedding in survivors was 37 days.
Case-control study

Schematic diagram of case-control study design

Grimes et al. Lancet 2002;359:431-34
Maternal and neonatal outcomes of pregnant women with COVID-19 pneumonia: a case–control study

Na Li, MD, Lefei Han, MPH, Min Peng, MD, Yuxia Lv, MM, Yin Ouyang, MM, Kui Liu, MM, Linli Yue, MM, Qiannan Li, MM, Guoqiang Sun, MM, Lin Chen, MD... Show more

Author Notes

Clinical Infectious Diseases, ciaa352, https://doi.org/10.1093/cid/ciaa352

- Pregnant women who were tested positive for SARS-CoV-2 were classified as laboratory confirmed case group
- The control group of pregnant women without pneumonia during hospital stay were randomly selected from the medical records
Cross-sectional study

Figure 1–22 Schematic representation of a cross-sectional study, conceptually and methodologically analogous to the case-based case-control study represented in Figure 1–19, except that instead of explicitly selecting cases and controls, it selects a sample of the entire population. Broken diagonal lines with arrows represent losses to follow-up. Cases are represented by “D” boxes.
COVID-19 Antibody Seroprevalence in Santa Clara County, California

Eran Bendavid\textsuperscript{1}, Bianca Mulaney\textsuperscript{2}, Neeraj Sood\textsuperscript{3}, Soleil Shah\textsuperscript{2}, Emilia Ling\textsuperscript{2}, Rebecca Bromley-Dulan\textsuperscript{2}, Cara Lai\textsuperscript{2}, Zoe Weissberg\textsuperscript{2}, Rodrigo Saavedra-Walker\textsuperscript{2}, Jim Tedrow\textsuperscript{4}, Dona Tversky\textsuperscript{6}, Andrew Bogan\textsuperscript{7}, Thomas Kupiec\textsuperscript{8}, Daniel Eichner\textsuperscript{9}, Ribhav Gupta\textsuperscript{10}, John P.A. Ioannidis\textsuperscript{1,10}, Jay Bhattacharya\textsuperscript{1}

Methods
On April 3-4, 2020, we tested county residents for antibodies to SARS-CoV-2 using a lateral flow immunoassay. Participants were recruited using Facebook ads targeting a sample of individuals living within the county by demographic and geographic characteristics. We estimate weights to adjust our sample to match the zip code, sex, and race/ethnicity distribution within the county. We report both the weighted and unweighted prevalence of antibodies to SARS-CoV-2. We also adjust for test performance characteristics by combining data from 16 independent samples obtained from manufacturer’s data, regulatory submissions, and independent evaluations: 13 samples for specificity (3,324 specimens) and 3 samples for sensitivity (157 specimens).

Results
The raw prevalence of antibodies to SARS-CoV-2 in our sample was 1.5% (exact binomial 95CI 1.1-2.0%). Test performance specificity in our data was 99.5% (95CI 99.2-99.7%) and sensitivity was 82.8% (95CI 76.0-88.4%). The unweighted prevalence adjusted for test performance characteristics was 1.2% (95CI 0.7-1.8%). After weighting for population demographics of Santa Clara County, the prevalence was 2.8% (95CI 1.3-4.7%), using bootstrap to estimate confidence bounds. These prevalence point estimates imply that 54,000 (95CI 25,000 to 91,000 using weighted prevalence; 23,000 with 95CI 14,000-35,000 using unweighted prevalence) people were infected in Santa Clara County by early April, many more than the approximately 1,000 confirmed cases at the time of the survey.

Ecologic Studies

• Explores correlations between aggregate (group level) exposure and outcomes
• Unit of analysis: not individual, but clusters (e.g. countries, counties, schools)
• Useful for generating hypothesis
• Prone to “ecological fallacy”
• Cannot adjust well for confounding due to lack of comparability (due to lack of data on all potential covariates)
Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study

Aaron Miller, Mac Josh Reandelar, Kimberly Fasciglione, Violeta Roumenova, Yan Li, Gonzalo H Otazu
doi: https://doi.org/10.1101/2020.03.24.20042937

Figure 1: Higher death rates were presented in countries that never implemented a universal BCG vaccination policy.
The role of vitamin D in the prevention of coronavirus disease 2019 infection and mortality

Petre Cristian Ilie¹ · Simina Stefanescu² · Lee Smith³

![Graph showing the relationship between mean vitamin D levels and COVID-19 outcomes.](image)

**Fig. 1** Mean vitamin D levels per country versus COVID-19 cases and mortality/1M population
Diagnostic accuracy studies

• Goal is to estimate the accuracy of the new test, compared to an established ‘gold standard’

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test +</strong></td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td><strong>Test -</strong></td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>
Improved Molecular Diagnosis of COVID-19 by the Novel, Highly Sensitive and Specific COVID-19-RdRp/Hel Real-Time Reverse Transcription-PCR Assay Validated In Vitro and with Clinical Specimens

Jasper Fuk-Woo Chan,†,b,c,d,e Cyril Chik-Yan Yip,† Kelvin Kai-Wang To,a,b,c,d Tommy Hing-Cheung Tang,†
Sally Cheuk-Ying Wong,‡ Kit-Hang Leung,§ Agnes Yim-Fong Fung,‖ Anthony Chin-Ki Ng,§ Zijiao Zou,‖ Hoi-Wah Tsoi,‖
Garnet Kwan-Yue Choi,‖ Anthony Raymond Tam,‖ Vincent Chi-Chung Cheng,‖ Kwok-Hung Chan,a,b,c,d Owen Tak-Yin Tsang,‖
Kwok-Yung Yuen,b,c,d,e
Outbreak investigations

• Could be:
  • Cohort study
  • Case-control study
  • Cross-sectional study

• Major steps:

- Detect a possible outbreak
- Define and find cases
- Generate hypotheses
  - Interviews
- Test the hypotheses
  - Analytic studies
  - Laboratory testing of samples
- If cases stop
  - Unsolved mystery
- Find link between food & illness
- No link found between food & illness
- Cases stop
- Control an outbreak
  - Recall product(s)
  - Remove source of contamination
  - Revise production process
- Find the point of contamination and source of the food

Source: CDC | Post-Dispatch
All of these “-emic”s! ...which is which?

- A disease is said to be endemic to a particular region or population when it is consistently present in that region / population.
- An outbreak is when the number of cases of a disease in a population increases above the normally expected (baseline) level.
- An outbreak can lead to an epidemic when the disease spreads to a large proportion of the population in a certain area or region.
- An epidemic becomes a pandemic when it spreads globally or across multiple countries or regions.

WHO, 2020 https://www.who.int/hac/about/definitions/en/
Epidemic Curve

Distribution of new cases of an infectious disease vs. time of reporting of cases (or onset of illness)

Helps to estimate incubation period, time of exposure, etc.

Epidemic curve of confirmed COVID-19 cases by WHO Region.
(Cases by date of reporting)

WHO, 2020
https://tinyurl.com/ycf8n9c
Attack Rate

Attack Rate = \frac{\text{# of new cases of disease during specified time interval}}{\text{# of people in the population at the start of this time interval}} \times 100

Also called incidence proportion, or risk

E.g: 100 new cases of COVID-19 in the last 24 hrs in a population of 500,000.
Attack rate = \frac{100}{500,000} \times 100 = 0.02

CDC, 2011
The Reproduction Number

Describes transmissibility of a communicable disease

**$R_0 = \text{Basic Reproduction Number:}**
The number of expected secondary cases generated by an infected individual in a completely susceptible population

**$R = \text{Effective Reproduction Number:}**
The number of expected secondary cases generated by an infected individual in a population including susceptible and non-susceptible individuals

Points of caution:
- Dependent on many factors
  - biological, environmental (e.g. population densities), etc
- Usually determined through modelling (covered later)
  - hence subject to uncertainty and to any assumptions the model may make

For COVID-19:
Accuracy of estimates of the $R_0$ (or $R$) is limited by our understanding of this new disease, e.g.:
- infectious period
- contribution of asymptomatic cases to transmission
- susceptibility of the overall population

E.g. If one person infected with disease X on average infects 3 others...

...$R_0 = 3$
(assuming susceptible population)

Now, if, say, $\frac{1}{3}$ of the population is immune...$R = 2$

Delamater et al. 2019  [https://dx.doi.org/10.3201/eid2501.171901](https://dx.doi.org/10.3201/eid2501.171901)
Epidemic doubling time

Time interval in which the cumulative incidence of a disease doubles.

Example:

If rate of testing and reporting remains the same

Muniz-Rodriguez, 2020
http://dx.doi.org/10.3201/eid2608.200219
Herd Immunity

• The resistance to the spread of a contagious disease within a population that results if a sufficiently high proportion of individuals are immune to the disease, especially through vaccination.

• When more people are immune, the effective reproductive number (R) is less than 1 and there is less transmission.

https://www.pbs.org/wgbh/nova/article/herd-immunity/
Research question drives the study design

“The question being asked determine the appropriate study architecture, strategy and tactics to be used - not tradition, authority, experts, paradigms or schools of thought.”

- Sackett, Wennberg 1997

- Good research starts with asking a clear, focused research question.
- How does one ask a focused research question?
How are these questions different?

• Does hydroxychloroquine improve outcomes among Covid-19 patients?

• In adult patients with lab-confirmed, severe Covid-19 illness, does daily hydroxychloroquine lead to higher survival rates as compared to placebo?
Architecture of a focused question: a 4-part review question

P - Who is the patient or what problem is being addressed?

I/E - What is the intervention or exposure?

C – What is the comparison group?

O - What is the outcome or endpoint?

Richardson et al. The well-built clinical question: a key to evidence-based decisions. ACP Journal Club 1995;A-12
Formulation of a therapy question

Is Remdesivir effective in treating Covid-19?

In adults with severe Covid-19, is intravenous Remdesivir effective in reducing mortality, as compared to placebo?

+ RCTs
Formulation of an etiology question

Is smoking a risk factor for Covid-19?

Are adults who smoke regularly at a greater risk (or protected) of developing Covid-19 as compared to those who do not smoke at all?

+ cohort & case-control studies
Type of questions dictate the design chosen

- Disease prevalence [cross-sectional survey]
- Etiology [cohort, case-control]
- Therapy [RCT]
- Prognosis [cohort or RCT]
- Harm [cohort, case-control]
- Diagnosis [cross-sectional, case-control, RCT]
- Economic [cost-effectiveness analysis, etc.]