How to write a research paper

- overview -

손희정
What is scientific paper?

• Paper organized description of hypothesis, data and conclusion to instruct the reader

• Realize your objective
  – to formulate and test hypotheses
  – To draw conclusion
  – not to collect data
Why you should care

• If you don’t publish it, you haven’t done it.
• The credit goes to the people who publishes first.
• People who don’t submit their work for peer review do crappy work
Why you should care less about writing a thesis or dissertation than a paper

• It is a lot more work
• You get less credit
• No one will ever read it
• You won’t be particularly proud of it
Why were papers rejected for publications?

• The study did not address an important scientific issue

• The study uncontrolled or inadequately controlled

Trisha Greenhalgh, BMJ 1997
Clinical study designs

• **Observational study**
  - Cohort study
  - Case-control study
  - Cross-sectional study

• **Experimental study**
  - Randomized controlled trial
Cross-sectional study design

Comparing people w factor and w/o factor **at the same time**.

- Snap shot
- Burden of illness and risk factors
- Service needs
- Hypothesis generation
Randomized Controlled Trial Issues

- Expensive
- Time consuming
- Questions
  - Comparisons of active treatments or all possible therapeutic options
  - Evaluation of rapidly changing interventions or technologies
- RCT provide the data in controlled setting rather than real world clinical setting
Why were papers rejected for publications?

• The statistical analysis was incorrect or inappropriate
How to Conduct Retrospective Cohort Study

- Decide dataset and period
- Download the whole data and freeze
- Data cleaning
- Developing data set and coding book
- Developing user guideline
- Developing data distribution and evaluation process
- Distributing data
Why is Data Cleaning Needed?

• Are all the data in place and accounted for, or are some of the data absent or missing?
  – Is there a pattern to the missing data?
• Are there any unusual or extreme responses present in the data set that may distort understanding of the phenomena?
Data Cleaning Tasks

- Identify outliers
- Fill in missing values
Data Cleaning-Outliers

• Detecting outliers
  – Graphically
  – Statistical tests

• Handling outliers
  – What caused the outlier?
  – What can be done to clean the data?
Detecting Outliers Graphically

- A boxplot can show outliers in continuous data.
- A histogram can show outliers in categorical data.
Detecting Outliers with Statistical Tests

- Percentiles
- Normal distribution
- Tukey

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Stem &amp; Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.5</td>
</tr>
<tr>
<td>3.00</td>
<td>2.001</td>
</tr>
<tr>
<td>3.00</td>
<td>2.555</td>
</tr>
<tr>
<td>7.00</td>
<td>3.0011222</td>
</tr>
<tr>
<td>9.00</td>
<td>3.555555588</td>
</tr>
<tr>
<td>20.00</td>
<td>4.00000000000000000122344</td>
</tr>
<tr>
<td>26.00</td>
<td>4.5555555555555555555555555788</td>
</tr>
<tr>
<td>29.00</td>
<td>5.000000000000000000000001111113</td>
</tr>
<tr>
<td>28.00</td>
<td>5.55555555555555555555555555888</td>
</tr>
</tbody>
</table>
Handling outliers

• Remove the outlier
  – Are the outliers data entry errors?
  – Are the outliers due to entries in the data set that don’t belong there?

• Keep the outliers
  – Are the outliers legitimate points that simply have extreme values?
Data Cleaning Tasks

- Identify outliers
- Fill in missing values
Missing Data

• Missing data are common
  – Inadequately handled in both epidemiological and experimental research
  – E.g., Wood et al. (2004) reviewed BMJ, JAMA, Lancet and NEJM.
    • 89% had partly missing outcome data
    • 46% performed complete case analysis

• Advanced methods have been developed to handling missing data
Why should you pay attention?

- Missing data contains potentially important information
- Ignoring or inappropriate handling data may lead to...
  - Biased estimates
  - Incorrect standard errors
  - Incorrect inferences/ results
Types of Missing Data

• Unit nonresponse: entire "unit" (e.g., individual) is missing
  – E.g., did not respond at all to follow up survey
  – Usually handled using nonresponse weighting

• Item nonresponse: individual items are missing
  – E.g., answered most of the questions, but a few blank
  – Usually handled using imputation approaches
Lots of Reasons for Missingness...

- Data entry errors - outside of range
- Administrative data with missing values
- Lost survey forms
- Individuals not wanting to disclose particular information
Missing Data Mechanism

• **Missing Completely at Random (MCAR):** Missingness is totally random; does not depend on anything
  – Probability of missing on variable Y is unrelated to the true value of Y or other variables or the patterns in the dataset
    • Ex. Water damage to paper forms prior to entry
  – No systematic differences between those with missing and observed values
  – Analyses using only complete cases will
    • have low power-small sample size
Missing Data Mechanism

- **Missing at random (MAR):** depends on observed data
  - Probability of missing on variable Y is unrelated to Y only after adjusting for one or more other variables in the dataset
    - Ex. Women more likely to respond than men
  - Use weighting or imputation
  - Probably the assumption made most frequently
  - If you ignore this, Bias & Efficiency
Ways of Handling Missing Data

- Improper ways of handling missing data
- Better ways of dealing with missing data
Inappropriate Ways of Handling Missing Data

- Complete case
- Single imputation
- Last observation carried forward
Complete-Case Analysis

- Restrict analyses to individuals with observed data
- Generally bad
  - Often results in maximum loss of cases
    - Decreased power and loss of representativeness
    - Especially in longitudinal setting
- Also called listwise deletion
Single Imputation

• Mean: e.g., mean age
  – Underestimate the variability in the imputed value

• Regression prediction
  – Impute mean within categories of observed covariates
  – $Y_i = \alpha + \beta X_i$

• Regression prediction plus error
  – Like regression prediction, but plus add random error
  – $Y_i = \alpha + \beta X_i + \epsilon_i$
Inappropriate Strategies

• Last observation carried forward
  – For longitudinal studies
    • Someone drops out of study, the last value observed for them is carried forward to later time points
  – But generally biased

<table>
<thead>
<tr>
<th>Year</th>
<th>BMI</th>
<th>LOCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1 (2000)</td>
<td>22 22</td>
<td></td>
</tr>
<tr>
<td>V2 (2003)</td>
<td>25 25</td>
<td></td>
</tr>
<tr>
<td>V3 (2006)</td>
<td>. 25</td>
<td></td>
</tr>
<tr>
<td>V4 (2009)</td>
<td>. 25</td>
<td></td>
</tr>
</tbody>
</table>
Appropriate Ways of Handling Missingness

• Weighting
• Multiple imputation

• Goal is not get correct predictions of missing values, goal is to obtain accurate parameter estimates for relationships of interest
  – reducing bias and inefficiency
Nonresponse Weighting

• Used to deal with attrition
• Generate the model predicting response given observed variable
• Weight respondents by their inverse probability if response
  – Weights the respondents up to represent the full sample
• Extreme weight
  – Check distribution of weights, trim outliers
Multiple Imputation

- Same idea as single imputation, but still fills in each missing value multiple times
- Creates multiple (e.g., 10) complete data sets
- Analyses then run separately on each datasets and results combined across datasets
  - Standard combing rules
Why were papers rejected for publications?

- The paper is so badly written that is incomprehensible
Writing papers is a skill

• Many papers are badly written
• Good writing is a skill you can learn
• It’s a skill that is worth learning:
  – You will get more brownie points (more papers accepted etc)
  – Your ideas will have more impact
  – You will have better ideas

How to write a great paper by Peyton Jones, Microsoft Research, University of Cambridge
Writing papers: model 1

Idea → Do research → Write paper
Writing papers: model 2

- Crystallizes what we don’t understand
- Opens the way to dialogue with others: reality check, critique, and collaboration
Do not be intimidated

**Fallacy**  You need to have a fantastic idea before you can write a paper. (Everyone else seems to.)

Write a paper, and give a talk, about **any idea**, no matter how insignificant it may seem to you.
Ideas

• Perspectives
  – Ideas originate from your collective knowledge of a topic
  – The more you read, the more ideas you will have
  – Without perspectives, it is difficult to generate ideas

Guidelines for writing successful research proposals in science by Samantha Joe, University of Georgia, 2006
Knowledge is power

• Know the literature
• Know who is funded to do similar work
• Clearly identify how your ideas are unique
• Allow time for your ideas to develop and mature
Papers communicate ideas

• Your goal: to infect the mind of your reader with your idea, like a virus
• Papers are far more durable than programs (think Mozart)

The greatest ideas are (literally) worthless if you keep them to yourself
Consider a bifircuated semi-lattice $D$, over a hyper-modulated signature $S$. Suppose $p_i$ is an element of $D$. Then we know for every such $p_i$ there is an epi-modulus $j$, such that $p_j < p_i$.

- Sounds impressive...but
- Sends readers to sleep
- In a paper you MUST provide the details, but FIRST convey the idea
Getting started

• Keep track of your ideas (idea book)
• Place your ideas in context (read)
• Formulate testable hypotheses
• Design experiments to test hypotheses
• Develop a conceptual model of the system
Presenting the idea

• Explain it as if you were speaking to someone using a whiteboard
• Once your reader has the intuition, they can follow the details (but not vice versa)
• Even if they skip the details, they still take away something valuable
Putting the reader first

• **Do not** recapitulate your personal journey of discovery. This route may be soaked with your blood, but that is not interesting to the reader.

• Instead, choose the most direct route to the idea.
Getting help

Get your paper read by as many friendly guinea pigs as possible

• Experts are good
• Non-experts are also very good
• Each reader can only read your paper for the first time once! So use them carefully
• Explain carefully what you want (“I got lost here” is much more important than “Jarva is mis-spelt”).
Pitching your idea

- E-mail
  - Introduce yourself
  - Pitch your idea
    - How it works fits into the broader scheme of supported projects
  - Pitch yourself
    - Why you are the person to do this work
Getting expert help

• Often they will respond with helpful critique

• They are likely to be your referees anyway, so getting their comments or criticism
Preparing

• Is organized, engaging and persuasive
  – Easy and fun to read
• Convince the reader that you
  – Understand the problem
  – Have a novel, appropriate and focused approach to address the problem
The process of writing
Structure of a research paper

**Experimental process**
- What did I do?
- What is the problem?
- How did I solve the problem?
- What did I find out?
- What does that mean?
- Who helped me out?
- Whose work did I use?

**Section of paper**
- Abstract
- Introduction
- Methods
- Results
- Discussion
- Acknowledgment
- Literature Cited
## Order of writing

<table>
<thead>
<tr>
<th>Section</th>
<th>Ideal</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>fifth</td>
<td>fifth</td>
</tr>
<tr>
<td>Introduction</td>
<td>first</td>
<td>third</td>
</tr>
<tr>
<td>Methods</td>
<td>second</td>
<td>first</td>
</tr>
<tr>
<td>Results</td>
<td>third</td>
<td>second</td>
</tr>
<tr>
<td>Discussion</td>
<td>fourth</td>
<td>fourth</td>
</tr>
<tr>
<td>Acknowledgment</td>
<td>last</td>
<td>last</td>
</tr>
<tr>
<td>Literature cited</td>
<td>always</td>
<td>always</td>
</tr>
</tbody>
</table>
What story are you telling?

• In science, the creative part is in the work, not the writing
• Scientific paper use very formal writing style that is very different than fun literature
• Write out a one page story of what you are trying to communicate
Your narrative flow

• Here is a problem
• It’s an interesting problem
• It’s an unsolved problem
• **Here is my idea**
• My idea works (details, data)
• Here’s how my idea compares to other people’s approaches
Make an outline

• Before writing any words or making any figures of your paper make an outline!
• There should be a minimum of one outline entry per paragraph
• Put your figures in your outline. Describe what the figure should communicate.

How to write a scientific research paper
Make an outline

• Your outline should tell you what you want to say and why you need to be said. What exactly are you trying to communicate?

• Put references (when you know them) in your outline
Make an outline

• NEVER start writing a section until you are satisfied with your outline and you have let your advisor look over and approve your outline
• Don’t underestimate how hard it is to write a good outline!
• The work is worth it.
Writing what you know about

• Don’t worry about having perfect spelling and grammar or getting all your references in the first time
  – If you don’t some number put in a place holder (i.e. xxx)
  – If you don’t know a reference put in a place holder (i.e. ref)
  – If you don’t like how the writing sounds, don’t worry, just make a comment. (i.e. rewrite this section)
Writing what you know about

• Write when you are fresh and alert. Edit your writing when you are tired or brain dead
  – It is very easy to get distracted
• Block out several hours a day away from distractions to write (i.e. library)
Writing what you know about

• If you get stuck refer back to your outline.
  – If it doesn’t help then stop writing and work on your outline instead

• This is not a paper you turn in for a grade-expect to have many revisions done by your advisor
Writing what you know about

• Don’t get upset or angry or take a revision personally
• Something is going to be changed? Put in a few spelling and grammar mistakes on purpose!
• Rename or renumber your documents every time you make a change
The process

• Start early. Very early.
  – Hastily-written papers get rejected.
  – Papers are like wine: they need time to mature

• Collaborate
Listening to your reviewers

Treat every review like gold dust
Be (truly) grateful for criticism as well as praise

This is really, really, really hard

But it’s really, really, really, really, really, really, really, really, really, really, really, really important
Listening to your reviewers

• Read every criticism as a positive suggestion for something you could explain more clearly
• DO NOT respond “you stupid person, I meant X”. Fix the paper so that X is apparent even to the stupidest reader.
• Thank them warmly. They have given up their time for you.
Related work

Fallacy  To make my work look good, I have to make other people’s work look bad
The truth: credit is not like money

Giving credit to others does not diminish the credit you get from your paper

- Warmly acknowledge people who have helped you
- Be generous to the competition. “In his inspiring paper [Foo98] Foogle shows.... We develop his foundation in the following ways...”
Credit is not like money

Failing to give credit to others can kill your paper

If you imply that an idea is yours, and the referee knows it is not, then either
- You don’t know that it’s an old idea (bad)
- You do know, but are pretending it’s yours (very bad)
Basic stuff

• Submit by the deadline
• Keep to the length restrictions
  – Do not narrow the margins
  – Do not use 6pt font
• Always use a spell checker
Presentation cosmetics

• Use concise, specific language

• Avoid universal terms (all, exact) and vague term (some, often)
Some points of style

• The word “this” must always be followed by a noun, so that its reference is explicit
  – Not: This is a fast reaction
  – But: This reaction is fast

• Describe experimental results uniformly in the past tense
  – Not: Addition of water gives product
  – But: Addition of water gave product
Some points of style

• Use the active voice whenever possible
  – Not: It was observed that the solution turned red
  – But: The solution turned red or We observed that the solution turned red

• Complete all comparisons
  – Not: The yield was higher using bromine
  – But: The yield was higher using bromine than chloride
Letter to the editor

• Why you are submitting the paper
• Why it is an important contribution
• Suggesting possible reviewers
Surviving peer review

• Wait
• Four levels of reviews
  – Accept
  – Accept pending minor revision
  – Accept pending major revision
  – Reject
• Famous (Cheville’s) rule of paper acceptance
  – 50% well known scientist, 20% slipped through, 20% really good paper, 10% affirmative action paper
Low-Fat Dietary Pattern and Risk of Colorectal Cancer
The Women's Health Initiative Randomized Controlled Dietary Modification Trial

CONCEPT: Observational studies and polyp recurrence trials are not conclusive regarding the effects of a low-fat dietary pattern on risk of colorectal cancer, necessitating a primary prevention trial.

OBJECTIVE: To evaluate the effects of a low-fat eating pattern on risk of colorectal cancer in postmenopausal women.

DESIGN, SETTING, AND PARTICIPANTS: The Women's Health Initiative Dietary Modification Trial, a randomized controlled trial conducted in 48,835 postmenopausal women aged 50 to 79 years recruited between 1993 and 1998 from 40 clinical centers throughout the United States.

INTERVENTIONS: Participants were randomly assigned to the dietary modification intervention (n=24,177; 40%) or the comparison group (n=24,658; 60%). The intensive behavioral modification program aimed to motivate and support reductions in dietary fat, to increase consumption of vegetables and fruits, and to increase grain servings by using group sessions, self-monitoring techniques, and other tailored and targeted strategies. Women in the comparison group continued their usual eating pattern.

MAIN OUTCOME MEASURE: Incident colorectal cancer incidence.

RESULTS: A total of 480 incident cases of invasive colorectal cancer occurred during a mean follow-up of 8.1 years (1.7 years). Intervention group participants significantly reduced their percentage of energy from fat by 10.7% more than did the comparison group. After 1 year, and this difference between groups was mostly maintained (6.1% at year 6). Statistically significant increases in vegetable, fruit, and grain servings were also noted. Despite these dietary changes, there was no evidence that the intervention reduced the risk of invasive colorectal cancer during the follow-up period. There were 201 women with invasive colorectal cancer (0.13% per year) in the intervention group and 279 (0.12% per year) in the comparison group (hazard ratio, 1.08; 95% confidence interval, 0.90-1.29). Secondary analyses suggested potential interactions with baseline aspirin use and combined estrogen-progestin use status (P=0.01 for each). Colorectal examination rates, although not protocol defined, were comparable between the intervention and comparison groups. Similar results were seen in analyses adjusting for adherence to the intervention.

CONCLUSION: In this study, a low-fat dietary pattern intervention did not reduce the risk of colorectal cancer in postmenopausal women during 8.1 years of follow-up.

Clinical Trials Registration ClinicalTrials.gov Identifier NCT00004411

JAMA. 2008;299:640-654

See also pp 629 and 655.