Supplementary Appendix 1. Examples of information typically shared by a participating site in a multi-center study

An example of information shared by a participating site with the analysis center in a multi-center study, when conducting pooled patient-level data analysis with individual covariates

Patient ID	Study treatment	Study outcome	Follow-up time	Covariate 1	Covariate 2	Covariate 3	Covariate 4	Covariate 5	…
001	1	0	312	0	0	0	1	1	…
002	1	0	40	1	0	0	2	0	…
003	1	0	365	1	0	0	2	0	…
004	1	1	200	2	1	1	1	0	…
005	0	1	20	3	1	2	3	0	…
006	0	1	15	3	1	0	2	1	…
007	0	0	14	1	0	3	2	1	…
008	0	0	145	0	0	1	3	0	…
009	0	0	355	2	1	2	3	0	--
…	…	…	…	…	…	…	…	…	…

In this dataset, each row represents a patient and each column represents a variable.

An example of information shared by a participating site with the analysis center in a multi-center study, when conducting pooled patient-level data analysis with confounder summary scores

Patient ID	Study treatment	Study outcome	Follow-up time	Propensity score
001	1	0	312	0.3244
002	1	0	40	0.1232
003	1	0	365	0.6578
004	1	1	200	0.1246
005	0	1	20	0.4569
006	0	1	15	0.0123
007	0	0	14	0.7086
008	0	0	145	0.5932
009	0	0	355	0.4959
…	…	…	…	…

In this dataset, each row represents a patient and each column represents a variable. Confounders are summarized into propensity scores, one of the most commonly used confounder summary scores.

An example of information shared by a participating site with the analysis center in a multi-center study, when conducting stratification analysis with confounder summary scores

Propensity score stratum	No. treated patients	No. untreated patients	No. outcome events among treated patients	No. outcome events among untreated patients
1	35	40	10	8
2	32	35	7	21
3	56	46	9	10
4	43	46	6	5

In this dataset, each row represents a stratum defined by propensity score, one of the most commonly used confounder summary scores. The cells provide the summary counts needed for the analysis.

An example of information shared by a participating site with the analysis center in a multi-center study, when conducting risk set-based analysis

Event time	Exposure status in patient who had the outcome	Exposure probability in risk set
2	0	0.34
15	1	0.21
35	0	0.11
45	0	0.05
47	1	0.67
79	1	0.88
111	1	0.10

In this dataset, each row represents an event time. The cells provide the information needed for the analysis.

An example of information shared by a participating site with the analysis center in a multi-center study, when conducting meta-analysis of site-specific effect estimates

Hazard ratio	Lower bound of 95% confidence interval	Upper bound of 95% confidence interval
0.68	0.45	1.02

The site only provides its effect estimate and information needed to calculate the site-specific weight (e.g., 95% confidence interval, standard error, or variance).

Supplementary Appendix 2. Fact sheet for the healthcare system leaders

Privacy-preserving analytic and data-sharing methods for clinical and patient-powered data networks

Please read this form carefully. It tells you important information about the research study. A research study is something you volunteer for; whether or not you take part is up to you. Someone on the research team will explain the study to you. Please ask them about anything you do not understand. You may ask to have this form read to you. If you do not understand what is in this form, do not participate in the study. You will be given a copy of the form to keep.

What is the purpose of this study?
Our ultimate goal is to help healthcare systems learn more, and more quickly, from routine healthcare information. This study is one step towards that goal. In this study, we hope to learn what people think about different ways of sharing and analyzing routine healthcare information.

What is routine healthcare information?
Healthcare organizations and insurance companies collect information on their patients and their patients' care. Examples include appointments, diagnoses, and medications.

Isn't healthcare information private?
Yes, this information is private. Organizations follow strict rules to protect patient privacy. In general they cannot share identifiable patient information without the patient's permission. There are exceptions, but they require special permission and review.

Can healthcare information be shared?
Yes, in some circumstances de-identified information can be shared.

What is de-identified data?
Data are de-identified when all information that could lead to an individual patient being identified is removed. This includes names, dates, medical record numbers and other personal information.

Why share healthcare information?
When information from large numbers of patients is combined, researchers can learn a lot. For example, some medication side effects are very rare, and are only discovered when very large numbers of patients take a medication.

Why does it matter what people think about data sharing?
We need to understand what people think about data sharing so that we develop safe, acceptable, and effective ways of learning from this data.

Why have I been asked to participate?
You have been asked to take part in this study because you are a health systems leader. We seek your input on how best to communicate how data is, or could be, shared and analyzed in multi-center research studies. Understanding your views will also help us to incorporate your preferences and suggestions into further improving these methods.

What would I be asked to do?
If you join the study you will be asked to review educational materials about data sharing. We will ask you questions about your understanding and opinions about data sharing. We may also ask you to complete some short questionnaires.

How long will the study last?
The study will last three years. Each meeting will last about 1-2 hours. We will also ask you to take part in:

No more than 3 in-person meetings
No more than 6 telephone discussions and
No more than 6 online meetings.

Is joining the study voluntary?
Yes, joining is completely voluntary. If you decide to join, you may change your mind later. You may quit the study at any time.

What happens if I decide to quit or not take part in this study?

If you choose not to be in the study, there will be no penalty to you, no impact on your employment, or loss of benefits. We will give you any new information during the course of this study that might change the way you feel about being in the study. If you would like to stop participating in the study you should let us know. If you decide to stop participating in the study, any information collected may still be used for this or future research.

How will the information I give be used and protected?
Study meetings will be recorded. Recordings will be transcribed. No personal information will be included on the transcripts. Your name will not be included on the transcript. Only approved study staff will read and analyze the transcripts. No identifying information will be included in any reports or publications. All paper forms will be kept in locked file cabinets and all information kept on computers will be password protected.

Are there possible benefits to taking part in this study?
You may not directly benefit from participation. However, you may learn something new about how your information collected by your health plan or healthcare provider may be used in research. You may enjoy participating and may feel that doing so contributes to scientific knowledge in general.

Are there any risks involved with taking part in this study?
Possible risks and discomforts from taking part in this study may include the potential loss of privacy or confidentiality resulting from unauthorized disclosure of information collected for this study, inconvenience in traveling to in-person interviews, and other potential risks that we currently cannot predict.

What if I am injured while taking part in this study?
This study involves meetings, interviews, and telephone calls only and so does not carry any risk of injury.

Will there be any costs to participating?
You will not be reimbursed for your own transportation cost traveling to and from the venue where the in-person interview will take place, or the phone or internet bills associated with the non in-person meetings.

Will I be paid for taking part in this study?
You will be paid $125.00 per hour for taking part in this study.

Who is funding the study?
This study is funded by The Patient-Centered Outcomes Research Institute (PCORI) [project number: ME-1403-11305].

Who do I contact about my rights as a research subject?
If you have any questions about your right as a research subject, you may contact the HPHC Institutional Review Board (IRB) at 1-800-807-6812.

Who is leading the study?
Darren Toh, ScD, is leading the study. Dr. Toh is an Associate Professor at Harvard Medical School and Harvard Pilgrim Health Care Institute. Dr. Toh would be happy to answer any questions. He can be reached at 617-509-9818 or Darren_Toh@harvardpilgrim.org. Additionally, the lead site investigator for Group Health Research Institute is Dr. David Arterburn. He can be reached at 206-287-4610 or arterburn.d@ghc.org.

Supplementary Appendix 3. Educational materials for the healthcare system leaders

Privacy-preserving analytic and data-sharing methods

Health Systems Leaders Stakeholder Meeting
Group Health Research Institute
Seattle, Washington
August 11, 2015

Goals of this session

- Describe different ways that health data could be shared and analyzed in multi-center studies
- Hear your questions, concerns, and advice

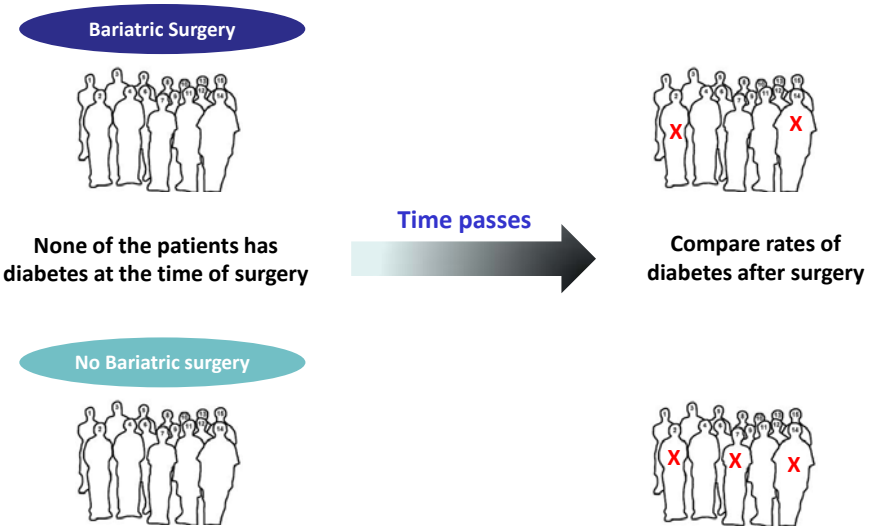
Case study

- Diabetes is a common and serious disease
- Diabetes is more common in overweight people than in average weight people
- Since we know that bariatric surgery helps reduce weight, does it also reduce the risk of diabetes?
- If bariatric surgery does reduce the risk of diabetes, does it have side effects that could outweigh this benefit?
- Do the benefits and risks differ by patient characteristics such as age and medical history?

Case study

- We can design a study to answer these questions

A hypothetical study



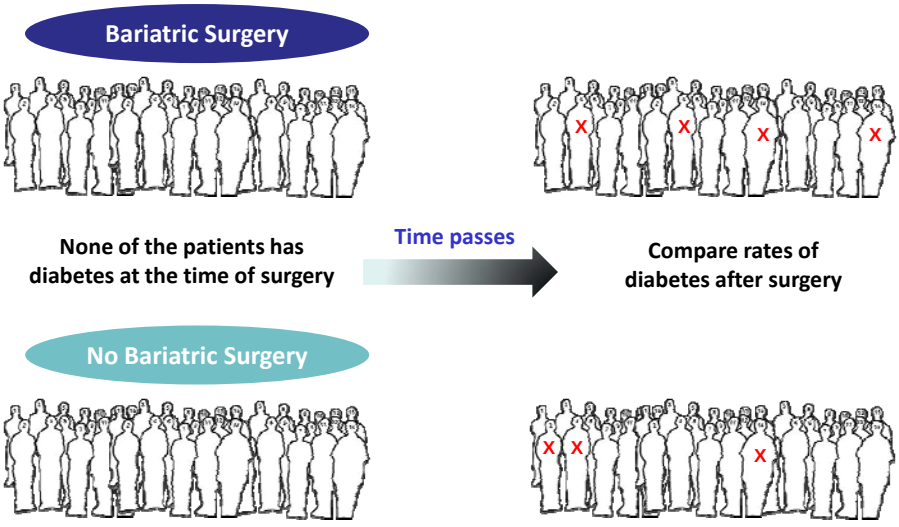
Patient-level info needed for the study

Patient ID	Bariatric Surgery	Diabetes Post Surgery	Follow-up Time	Age Group	Sex	Race	BMI Category	Heart Disease	…
001	1	0	312	0	M	0	1	1	…
002	1	0	40	1	M	0	2	0	…
003	1	0	365	1	F	0	2	0	…
004	1	0	200	2	F	1	1	0	…
005	0	1	20	3	F	2	3	0	…
006	0	1	15	3	M	0	2	1	…
007	0	0	14	1	M	3	2	1	…
008	0	0	145	0	F	1	3	0	…
009	0	0	355	2	M	2	3	0	--
010	…	…	…	…	…	…	…	…	…

A hypothetical study

- If *one* site participates in this study, we may only be able to include a small number of patients

A hypothetical study – Now with more sites



Patient-level info needed for the multi-site study

Site A

Patient ID	Bariatric Surgery	Diabetes Post surgery	Follow-up Time	Age Group	Sex	Race	BMI	Heart Disease	…
001	1	0	312	0	M	0	1	1	…
002	1	0	40	1	M	0	2	0	…
003	1	0	365	1	F	0	2	0	…
004	1	0	200	2	F	1	1	0	…
005	0	1	2	3	F	0	3	0	…
006	0	1	15	3	M	0	1	1	…
007	0	0	4	1	M	1	1	1	…
008	0	0	145	0	F	1	3	0	…
009	…	…	…	…	…	…	…	…	…

Site B

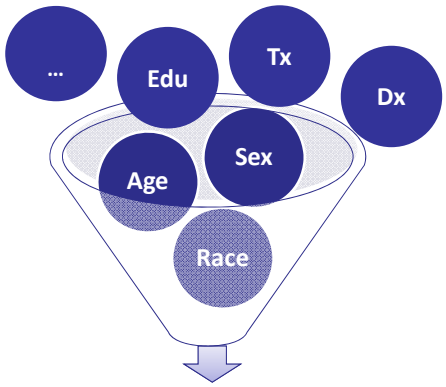
Patient ID	Bariatric Surgery	Diabetes Post surgery	Follow-up Time	Age Group	Sex	Race	BMI	Heart Disease	…
001	0	1	35	1	F	1	3	0	…
002	0	1	213	2	M	1	1	1	…
003	0	1	453	2	M	0	4	1	…
004	0	0	58	3	M	0	3	1	…
005	1	0	31	3	M	0	3	0	…
006	1	0	56	1	F	1	2	0	…
007	1	0	123	1	F	1	1	1	…
008	1	0	546	0	M	0	3	0	…
009	…	…	…	…	…	…	…	…	…

- Patient-level info can generally be *de-identified* so that sensitive patient info is not shared
- But even so, concerns about patient privacy or data security may still persist
- Sometimes it is not possible to share patient-level info due to these concerns or other reasons

Question

- Do we have other ways to share data?

We can collapse multiple variables into a summary score



Propensity Score or
Disease Risk Score

When we use these summary scores

Before summarization

Patient ID	Bariatric Surgery	Diabetes Post Surgery	Follow-up Time	Age	Sex	Race	BMI Category	Heart Disease	…..
001	1	0	312	0	1	0	1	1	0

Collapsing individual variables into a summary score makes it less likely that a patient be identified by their unique characteristics and medical history

After summarization

Patient ID	Bariatric Surgery	Diabetes Post Surgery	Follow-up Time	Age	Sex	Race	BMI Category	Heart Disease	…..
001	1	0	312	0	1	0	1	1	0

Before summarization

Patient ID	Bariatric Surgery	Diabetes Post Surgery	Follow-up Time	Age Group	Sex	Race	BMI Category	Heart Disease	…
001	1	0	312	0	M	0	1	1	…
002	1	0	40	1	M	0	2	0	…
003	1	0	365	1	F	0	2	0	…
004	1	0	200	2	F	1	1	0	…
005	0	1	20	3	F	2	3	0	…
006	0	1	15	3	M	0	2	1	…
007	0	0	14	1	M	3	2	1	…
008	0	0	145	0	F	1	3	0	…
009	0	0	355	2	M	2	3	0	--
010	…	…	…	…	…	…	…	…	…

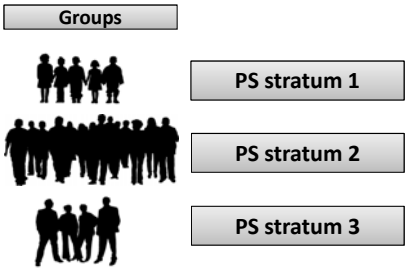
After summarization

Patient ID	Bariatric Surgery	Diabetes Post Surgery	Follow-up Time	PS
001	1	0	312	0.34
002	1	0	40	0.32
003	1	0	365	0.12
004	1	0	200	0.56
005	0	1	20	0.33
006	0	1	15	0.78
007	0	0	14	0.21
008	0	0	145	0.43
009	0	0	355	0.63
010	…	…	…	…

Methods that share summary-level info

- There are several other ways to share info in multi-center studies
- Some of these methods do not even require sharing of patient-level info
- Each of these methods has its pros and cons that are beyond the scope of our discussion today
- The next two slides show two of these methods that only require sharing of summary-level dataset

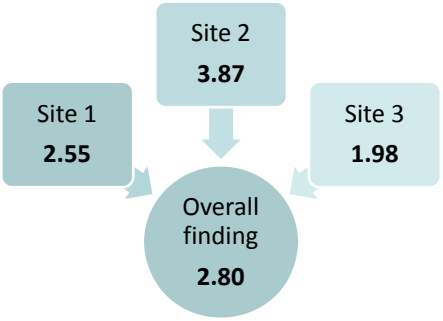
Methods that share summary-level info #1



- Each patient is placed into a group with similar patients
- We then use statistical techniques to combine across strata

PS stratum	Patients with bariatric surgery	Patients without bariatric surgery	Diabetes in patients with bariatric surgery	Diabetes in patients without bariatric surgery
1	35	40	10	8
2	32	35	7	21
3	56	46	9	10

Methods that share summary-level info #2



- Each site estimates the bariatric surgery-diabetes association individually
- The effect estimate is shared with the lead study team
- The lead study team aggregates the site-specific estimates

Patient-level vs. summary-level info

	Patient-level dataset	Summary-level dataset
Provides better protection for patient privacy		✓
Provides better protection for data security		✓

Patient-level vs. summary-level info

	Patient-level dataset	Summary-level dataset
Provides better protection for patient privacy		✓
Provides better protection for data security		✓
Allows analysis to be done more easily	✓	
Allows more analysis to be done	✓	

Discussion questions

- What questions do you have about patient-level vs. summary-level data-sharing methods?
- What are you most concerned about with respect to data sharing in research?
- Would the summary-level data-sharing methods described here help mitigate your concerns?
- What are your concerns about these summary-level data-sharing methods?

About this project

- **Who is funding the study?**
  The Patient-Centered Outcomes Research Institute (PCORI)

- **Who do I contact about my rights as a research subject?**
  HPHC Institutional Review Board (IRB) at 1-800-807-6812

- **Who is leading the study?**
  Darren Toh, ScD
  Associate Professor
  Harvard Medical School / Harvard Pilgrim Health Care Institute
  617-509-9818
  Darren_Toh@harvardpilgrim.org

Supplementary Appendix 4. Interview guides for the healthcare system leaders

1. **Can you briefly describe your roles and responsibilities in your organization?**

   Prompt if not clear:

   Do you have responsibility for research activities, or decision making around those activities?

   If yes; and not already clear, *prompt further for role in this context*.

   As you know, our goal is to understand your views on data sharing in the context of multi-site research studies. Let me give you an example of the sort of study we're particularly interested in and the types of data that might be shared.

   Insert brief example, end with sharing of individual-level data. (Refer to slides 1-9 of the educational materials)

2. **Has your site been involved in studies of this sort, that is, studies which have involved sharing data with investigators not at your site?**

   If organization has participated, prompt:

   Can you say more about that?

   Further probe (if needed):

   What types of data have been shared – for example, patient-level or summary-level information?

3. **Were you involved in those studies in any way – for instance, as an investigator, as someone who made decisions about whether to participate, or some other way?**

   If yes, probe to explore involvement:

   How were you involved?

4. **Do you feel that your organization has benefited from participating in studies that have involved data sharing?**

   Can you say more about that?

5. **Did you or others in your organization have concerns or reservations about sharing such information in these studies?**

Prompt, if yes:

How were those concerns addressed?

Was the study protocol or data sharing approach changed to address those concerns?

If yes:

Can tell me about those changes?

Were there any problems caused by sharing data? For instance, has anything ever gone wrong when your organization has participated in studies involving data sharing?

6. Have you encountered any challenges or barriers in participating in these sorts of studies?

If yes:

Please say more about those [challenges/barriers].

How have you responded to those [challenges/barriers]?

7. As you know from the information we sent you in preparation for this interview, we are particularly interested in your views on privacy-protecting methods for data sharing.

Refer to slides 10-17 of the educational materials, and then point out – briefly – the difference between sending individual-level vs. summary-level data.

Do you have any questions about any of these?

Are you familiar with any of these privacy-protecting methods?

8. Do you know whether your organization has ever shared data using one of these methods?

Prompt, if yes:

Can you tell me about that experience (or those experiences)?

9. Do you see a plus side to using privacy-protecting methods?

Or, if we give them information on what we consider the plus side, ask them to comment on those – do they see those as benefits?

Alternative 9a: Do you see the added protection afforded by these methods as important?

Prompt:

Can you say more about that?

Do you see any other benefits to using these approaches?

10. Do you see a down side to using privacy-protecting methods?

Or, if we give them information on what we consider the limitations/downside, ask them to comment on those.

Do you see any other limitations/downsides to using these approaches?

11. Do you have any final comments about any of the issues we've been talking about today?

Supplementary Appendix 5. Coding framework, including codes, themes, and subthemes

Themes Subthemes	Codes
Potential for harm, risks, concerns	
Potential harms	• Implied or explicit refer to potential harm to reputation, to commercial interests, proprietary concerns of organization • Implied or explicit refer to potential harm to reputation of provider • Potential for researcher to lose academic or research advantage • Potential harm to patients, includes reference to breach of confidentiality; unauthorized disclosure of protected health information; release of sensitive information like mental health issues, alcohol, HIV status; potential that data could be use against patient • Explicit statements that risk is minimal or absent
Experiences related to harms	• Personal experience with data sharing – any direct experience what harm or reference to absence of harm in personal experience • "Near miss" with harm • Reference to awareness of data breach, but outside of personal experience (e.g., VA data breach)
Concerns about data sharing	• Lose control over data when shared • General questions, reference to issues that one worries/wonders about • Concerns about others profiting from patient data; commercial use
Questions and concerns related to safeguards, data security	• Does requestor have the appropriate systems in place to protect data • How will protections, processes be confirmed • Who will de-identify the data • Who will have access to the data • How will data be transferred • Need to confirm that data are de-identified, not re-identifiable
Steps taken to minimize risk associated with data sharing	• Share only with legitimate or known investigators • Require local investigator to be involved • Data use agreements • Training (e.g., of programmers, biostatisticians, people who deal with the data) especially around data sharing and storage
Patient preferences related to data sharing	• Desire for transparency; awareness that sharing is occurring • Desire for education about how data is being used
Willingness to share data	
Factors affecting willingness	• More motivated to share when answer could benefit patient, improve patient outcomes • More motivated to share if research question focused on care delivery • Some data more sensitive, less likely to share • Patient desire to select what information will be shared, to opt in or out • Share only data related to study question • Type of organization, requestor influences willingness to share • Direct personal experience with data and/or research provides insight, reassurance related to lack of harm • Trust in those data is to be shared with (organizations, researchers at other sites, within organization) influences willingness to share
Influence of trust and relationships	• Influence of requestor's role, organization, intended use on trust • Trust influenced by experience, history

	• Trust develops over time, requires ongoing relationship • Trust higher when assurances that those receiving have experience, understand issues and need for safeguards
Benefits and value of data sharing, research	
Benefits of data sharing	• Data sharing generally desirable, beneficial • More data/multi-site research results in more generalizable findings, increases validity • Large datasets helpful for studying rare outcomes • Data sharing may contribute to better care
Value of research	• Value studies/recognize need for studies that answer an important question, improve patient care and/or patient outcomes • Value good science, credible results • Patients want their data to be helpful/useful
Costs	
	• Data sharing requires resources, need for validation, time for people to explain data • Resources are limited • Commitment of resources to data sharing, creating aggregate data is an opportunity cost • Cannot participate in study if funding does not cover costs • Reference to patient desiring compensation, appreciation, results
Views on sharing of individual vs. aggregate data/ granularity of data to be shared	
Individual-level data preferred	• Individual-level data is preferable (in general, not further specified) • Individual-level datasets perceived as less expensive to create, more multipurpose; greater ability to answer research questions, especially ones that emerge throughout a study • Aggregate data as a compromise (less valuable than individual data but better than none) • Concerns about how missing data are handled with aggregate data • Inability to "get dirty" with aggregate data; individual-level data allows more interaction with data; exploration of nuances • Some approaches (including privacy-protecting methods) requires greater technical and programming expertise
Aggregate data preferred	• Aggregate data forces researchers to specify research questions, variables, in advance (voiced as an advantage) • Aggregate data may facilitate studies that are otherwise challenging to conduct due to patient concern; privacy/data sharing concerns
Privacy-protecting methods – pros and cons	• Generally preferable, not specified further • Offer greater privacy protection, data security • May be more acceptable to IRB • May be more acceptable to patients • Belief there is a need for privacy-protecting methods • Belief there is not a need; current approaches to data sharing sufficient • Sufficiency of current approaches would change if somebody "made a big mistake"; or if other sites started to change their approach
Comments related to increasing acceptance of privacy-protecting methods	
	• Provide more examples, variety of examples, greater exposure to method • Demonstrate equivalence of results across analyses using individual-level data vs. privacy-protecting methods • Identify ways to implement PPM cheaper/faster/more efficiently