TC IRB Researcher Guidance on Secondary Analysis of Online Data and IRB Review

Regulatory Context

45 CFR 46.104(d)(4)(i) permits exemption for: *Secondary research uses of identifiable private information or identifiable biospecimens, if the identifiable private information or identifiable biospecimens are publicly available.*

This category is particularly relevant for digital and online research where investigators analyze **publicly available online content generated independent of the current research**, for example, social-media posts, online forums, blogs, or digital archives—without any new elicitation of data or interaction with individuals.

TC IRB Interpretive Approach

Although 45 CFR 46.104(d)(4)(i) refers to "identifiable private information," Teachers College IRB reviews secondary analyses of **publicly available identifiable online data** under Exempt Category 4(i), rather than as non-human subjects research (NHSR). This approach allows the IRB to confirm public accessibility, assess identifiability and re-identification risk, and ensure that use of public identifiers does not introduce ethical or reputational harm.

Core Concept

Category 4(i) applies when **both conditions** are met:

1. **The data is generated independent of the current research**, where the investigator does not interact with individuals or elicit new information to create the dataset**, and**
2. **The data are publicly available** to any member of the general public without restriction, login, or membership gatekeeping.

For TC IRB purposes, "private" is evaluated based on identifiability and reasonable expectations of privacy, not solely on whether content is technically accessible to the public.

Use of identifiable public information (e.g., usernames, posts, or profile photos) under this category is allowed **if** those identifiers are themselves part of the public domain. Data may be dynamically retrieved (e.g., via APIs or automated tools) and still qualify as secondary use, provided the content was generated independently of the investigator and no interaction or elicitation occurs.

In contrast, **NHSR** (non-human subjects research) applies only when the researcher is *not using identifiable information at all.*

Identifiability and Re-Identification Risk

Identifiers to consider:

- Usernames or handles.
- Profile photos or avatars.
- Specific timestamps. Specific timestamps can be used to:
    - Match a post, message, or activity to an identifiable person

Teachers College, Columbia University
Institutional Review Board
525 W. 120th St. Box 151, New York, NY 10027
212-678-4105 | IRB@tc.edu | RH 13 | tc.edu/IRB

  - Cross reference with publicly available content
  - Search exact moments in forums, social media, email threads, or logs

  When combined with exact quotes, usernames, handles, profile photos or avatars, a timestamp can allow someone to say: *"I know exactly who posted this, when, and where."* That makes the individual identifiable, even if names are removed.
- Exact quotes (searchable online).

How Re-Identification Can Occur

- **Searchable verbatim quotes**: Copying a long or distinctive quote can allow anyone to paste it into a search engine and locate the original post and user profile.
- **Unique usernames**: Handles like @RareDiseaseDad or @SingleTeacherInHarlem are uncommon enough to be directly traceable.
- **Image clues**: Avatars, profile photos, or background images may reveal a person's identity via reverse-image search or recognition within a community.
- **Contextual triangulation**: Details such as workplace, school program, neighborhood, or medical condition can identify someone when combined.
- **Niche or small communities**: In groups with only a handful of posters, even anonymized quotes may implicitly reveal who said what.

Best Practices

- Use paraphrasing when direct quotes enable re-identification.
- Avoid screenshots unless fully justified and anonymized (e.g., blurring faces, cropping out any identifiable information).
- Remove or code identifiers when possible.

Distinguishing NHSR vs. Exempt 4(i)

Criterion	NHSR	Exempt Category 4(i)
Data accessibility	Public and non-identifiable.	Publicly available, may include identifiers.
Identifiability	None — researcher ensures no identifiable or private information is collected. NHSR **does not apply** if identifiers (e.g., usernames, handles, timestamps, post IDs) are retained at any stage of data scraping, cleaning, validation, or analysis, even if removed prior to publication.	May retain identifiers if they are already publicly available (e.g., a public username or blog name).
Data collection status	Investigator collects or observes general patterns, not individual data.	Data are generated independent of the current research and are accessible to the general public.
Example data source	Aggregated counts of post frequencies.	Public Facebook pages, public Reddit threads,

Teachers College, Columbia University
Institutional Review Board
525 W. 120th St. Box 151, New York, NY 10027
212-678-4105 | IRB@tc.edu | RH 13 | tc.edu/IRB

		open-access blogs, public YouTube comments.
IRB oversight	Not required (Research Determination Form submission required for formal NHSR letter).	Required exemption determination to confirm public availability, no sensitive content and minimal ethical risk.

When Category 4(i) Applies

- The dataset is **publicly viewable by anyone** without login or group membership (e.g., public Reddit subreddits, open Facebook pages, X/Twitter public accounts).
- The data were generated independent of the current research; no researcher interaction or elicitation created or influenced the content.
- The researcher's use of identifiable elements (e.g., username, post ID, timestamp) is **necessary to contextualize findings** or **interpret discourse** but does not increase participant risk.
- Researchers should also clearly document in their IRB submission why identifiable information is technically unavoidable, how the data qualify as publicly available, and what steps will be taken to limit downstream use of identifiers. This includes describing data storage protections, plans for deidentification, and whether direct quotations, screenshots, or usernames will be altered or paraphrased in dissemination materials.
- The **information itself was intended for public consumption**, not shared in a context implying privacy or confidentiality.

When Category 4(i) Does Not Apply

- The content is located behind a **login wall** (e.g., Facebook "private" or "members-only" groups, Discord servers, closed forums).
- Users must **request to join** or be **approved by a moderator** to view or post content.
- The data contain **sensitive personal disclosures** (e.g., health status, trauma narratives) where users have a reasonable expectation of privacy—even if technically accessible.
- The investigator **elicits new data** or **interacts** with participants to obtain clarification or permission.
- The dataset includes **non-public metadata** (e.g., scraped user IP addresses or hidden account details).

If any of these apply, the study may be reviewed under a different exempt category or may require **expedited/full IRB review**.

Examples: Category 4(i) in Digital Research

Example	Determination	Rationale / IRB Consideration
Researcher downloads 5000 public posts using a hashtag (#VapeAwareness) via Twitter API, retaining usernames and text.	**Exempt 4(i)**	Posts are publicly accessible; identifiers are public. Minimal risk.

Teachers College, Columbia University
Institutional Review Board
525 W. 120th St. Box 151, New York, NY 10027
212-678-4105 | IRB@tc.edu | RH 13 | tc.edu/IRB

Investigator analyzes comments on public YouTube channels discussing fitness influencers.	**Exempt 4(i)**	Publicly visible content, no login or private group membership required.
PI examines Reddit discussions from r/AskDocs (public subreddit) using text mining; usernames included in dataset.	**Exempt 4(i)**	Publicly available; identifiers are part of public domain.
Researcher reviews archived blogs about breast cancer experiences from open-access platforms.	**Exempt 4(i)**	Public blogs; posts intended for general readership.
Investigator uses a de-identified Reddit dataset already published on Kaggle.	**NHSR**	Data are de-identified and not reasonably re-identifiable; no human subjects.
Researcher tracks frequency of public hashtags (#ClimateChange) over time without storing tweet text or usernames. Data are aggregated counts, no identifiers, no interaction.	**NHSR**	Data are aggregated counts, no identifiers, no interaction.
Investigator analyzes an anonymized corpus of Reddit posts already published by a data-science repository. Data not generated through investigator interaction or intervention, and de-identified; no link to living individuals.	**NHSR**	Data generated independent of the current research and de-identified; no link to living individuals.
Researcher examines only metadata (publication date, word count) of public blogs. No identifiable information analyzed.	**NHSR**	No identifiable information analyzed.
Researcher extracts posts from a closed Facebook group for veterans that requires membership approval.	**Not Exempt 4(i); another exempt category, OR expedited review, may apply**	Closed group = not publicly available; privacy expected.

Documentation Requirements for NHSR

Investigators must answer **"no"** to each of these before self-classifying as NHSR:

1. **Interaction:** Am I directly communicating with or influencing any individuals online?
2. **Private Information:** Am I collecting or recording data that could identify a living person?
3. **Expectation of Privacy:** Could users reasonably expect privacy in the space I'm observing?
4. **Sensitive Content:** Would disclosure of this data cause harm or embarrassment?
5. **Data Origin:** Was this content created for a public audience?
6. **Data Handling:** Will I retain or share any identifiers, even temporarily?
7. **Re-identification Risk:** Could someone reconstruct identities from my dataset?

Teachers College, Columbia University
Institutional Review Board
525 W. 120th St. Box 151, New York, NY 10027
212-678-4105 | IRB@tc.edu | RH 13 | tc.edu/IRB

If the answer to any question is **yes**, the project likely moves from **NHSR → Exempt 4(i)** or another review category.

Documentation Requirements for Category 4(i)

When claiming Exempt 4(i), researchers must include in the IRB application:

1. **Description of data accessibility:**
   - Demonstrate that anyone (logged out, non-member) can view the data source.
   - Screenshots or links showing open access are ideal.
2. **Statement on public identifiability:**
   - Explain that any identifiers retained (e.g., usernames) are public and necessary for data integrity.
   - Confirm no effort to link to non-public data.
3. **Assurance of minimal risk:**
   - Describe why inclusion of public identifiers poses no harm or reputational risk.
   - Avoid quoting or reposting sensitive disclosures out of context.
4. **Data management plan:**
   - Specify how identifiers will be stored, coded, or removed for publication.
   - Indicate whether data will be shared, archived, or restricted post-study.

IRB Review Focus Points

The IRB's evaluation of Exempt 4(i) typically centers on:

- **Is the data *truly* publicly available?**
  "Login required" or "private group" = not public.
- **Does the dataset include sensitive content?**
  If so, privacy expectations may override technical availability.
- **Could re-publication cause harm or embarrassment?**
  If yes, even public identifiers may need redaction or paraphrasing.

When identifiable public data are involved, TC IRB generally prefers Exempt 4(i) review to document this assessment, even when the investigator believes the activity could qualify as NHSR.

If there is *any ambiguity* about privacy expectations, the IRB may reclassify the study under **other review category(ies)** for additional scrutiny. *Note that although the category may differ, the mechanism of submission would be the same for either category (e.g., via Mentor IRB).*

Legal and Contractual Considerations (when applicable)

Issues related to platform terms of service, data use agreements, licensing, or other legal or contractual requirements fall outside the scope of IRB ethical review and should be consulted with the appropriate institutional offices (e.g., Office of General Counsel). Such reviews are handled separately and independently from IRB review and do not replace the requirement for IRB determination when applicable.

Ethical Reminder

Public availability ≠ ethical free-for-all.

Teachers College, Columbia University
Institutional Review Board
525 W. 120th St. Box 151, New York, NY 10027
212-678-4105 | IRB@tc.edu | RH 13 | tc.edu/IRB

Even in Category 4(i), researchers must exercise discretion:

- Avoid using **direct quotes** that could enable re-identification through search.
- Do not include **images or avatars** unless deidentification procedures are in place (e.g., face blurring, removal of any identifiers included).
- Acknowledge that participants may not anticipate their content being studied, even if posted publicly.

The guiding principle remains: *If the data are public, use is permissible — but if their use could still cause harm, obtain IRB confirmation and mitigate that risk.*

Sample IRB Language

Exempt 4(i)

"The dataset consists of publicly available online content that does not require login or membership to access. Identifiers are publicly visible and will be handled in accordance with privacy best practices."

NHSR

"The dataset contains no identifiable information and consists solely of de-identified online content generated independent of the current research. No interaction with individuals will occur."

Responsibilities

PI

- Accurately classify recruitment platforms (e.g., distinguish between public, semi-public (restricted access), and private groups).
- Secure permission from group administrators when posting in private or closed spaces.
- Document access rights or permissions clearly in the IRB application.
- Ensure recruitment respects the privacy expectations of group members and complies with site terms of use.

IRB

- Evaluate whether the proposed recruitment plan aligns with ethical standards and respects participant privacy.
- Verify internal consistency in the submission (e.g., claiming a group is public, but show it's private).
- Request clarification or corrections when discrepancies are noticed—but not proactively investigate or fact-check each claim unless something flags it as inconsistent.

Resources

- **45 CFR 46 – The Common Rule**
  U.S. Department of Health and Human Services regulations for the protection of human research subjects.
  https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/
- **SACHRP (2013).**
  Considerations and Recommendations Concerning Internet Research and Human Subjects Research Regulations
  Final report by the Secretary's Advisory Committee on Human Research Protections (SACHRP).

Teachers College, Columbia University
Institutional Review Board
525 W. 120th St. Box 151, New York, NY 10027
212-678-4105 | IRB@tc.edu | RH 13 | tc.edu/IRB

- **Advarra IRB.**
  Differentiating "Public" and "Private" Internet Spaces in IRB Review
  Practical guidance on privacy expectations and recruitment ethics in digital environments.
- **College of Charleston IRB.**
  Guidance on Research Using Social Networking Sites
  Clarifies researcher responsibilities and platform-specific expectations.
- **Facebook Group Privacy Settings**
  https://www.facebook.com/help/220336891328465
- **Instagram Privacy and Visibility**
  https://help.instagram.com/517073653436611
- **X (formerly Twitter) Protected Tweets**
  https://help.x.com/en/safety-and-security/how-to-make-x-private-and-public
- **LinkedIn Group Privacy Descriptions**
  https://www.linkedin.com/help/linkedin/answer/a548061
- **Reddit Community Types**
  https://support.reddithelp.com/hc/en-us/articles/360060416112

Appendix A
Submission Pathway Summary for Secondary Analysis of Online Data

After reviewing the guidance above, investigators should use the summary below to determine the appropriate submission pathway in Mentor IRB (e.g., via the **NHSR determination instance** or the **IRB instance in Mentor IRB**).

Decision Summary

If your project involves…	Submit As
Publicly available online data with no interaction or elicitation of data from individuals at any stage of data handling, and no identifiable information retained (e.g., aggregate or fully de-identified public data)	***NHSR determination**
Publicly available online data with no interaction or elicitation of data from individuals at any stage of data handling, with identifiers retained (e.g., usernames, handles, timestamps), and no sensitive content	****IRB application – Exempt Category 4(i)**
Data behind login or restricted access, sensitive content, reasonable expectation of privacy, or risk of harm	****IRB application – Expedited or Full Review**

*Please refer to the ***18_TC IRB Walkthrough for Submitting a Research Determination Form in Mentor IRB_2025_TC IRB.pdf*** document for steps by steps guidance on how to submit for an NHSR determination.

**Please refer to Submitting a New IRB Protocol and the Training & Education sections of the IRB website for steps by step guidance on how to submit a regular IRB application.

Teachers College, Columbia University
Institutional Review Board
525 W. 120th St. Box 151, New York, NY 10027
212-678-4105 | IRB@tc.edu | RH 13 | tc.edu/IRB