2022 MIDI Task Group Meeting Notes

January 11, 2022 Meeting

Interim Report Best Practices And Recommendations Extract as of 20220107

- Fred asked if the report should be focused on the US given that the details can differ geographically.
- Data created from European persons may not satisfy GDPR.
- We should highlight when this is true, along with caveats and any possible workarounds.
- Fred likes the ideas of universal guidelines to recommend to the EU.
- We will share the report with international colleagues once it the report is fleshed out.
- California regulations exclude healthcare data.
- Is it fair to focus on ethical and moral concerns as well as the legal concerns? We're trying to reduce the actual re-id risk and harm.
- So far we're focused on DICOM images.
- Kathy: Say anything about raw data signals?
- Wyatt: DICOM SR objects and embedded PDFs? Non-image objects, RT plans.
- Need a more precise definition for unrecognized. It is the opposite of "what is known to be safe."
- Specify what constitutes due diligence as you conduct your risk analysis. Can't help the unknown unknowns.
- Make the definition of collection clear. Collection doesn't communicate "version."
- "Release" not as good as "collection."
- "Indirect" and "direct" identifiers, sensitive information–a disease that may make someone discriminate against you or function as an indirect identifier.
- Ideally, you'd want to quantify the percentage of data elements you will be retaining.
- The paper will highlight the uncertainty.
- Steve: Address optional attributes as well.
- Calibration information can identify the machine used.
- Consistency of acquisition protocols.
- Need to consider and determine which options to the profile are selected.
- Part 15 and best practices are different.
- Only got through item 6 in the Summary of Best Practices. Will pick this up at the next meeting. To save time, team members can send David their comments in writing.

Action: Review the Interim Report and email David Clunie your comments.

February 8, 2022 Meeting

Interim Report Best Practices And Recommendations Extract as of 20220208

ITEM 6

- A paper in BMJ and Trials [Hrynaszkiewicz et al] in which the editor said it's okay to keep three patient characteristics, but more than that requires expert statistical analysis of re-identification risk.
- For this report, David C. is leaning towards saying that if any characteristics are retained, a statistical analysis should be performed, based on evaluation in IOM Report Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk that suggest no empirical basis for rule of two or three quasi-identifiers (Appendix B [El Emam & Malin]).
- Fred asks what the risk threshold we are comfortable with. David says we need the analysis first, then compare that against the risk threshold.
- David recommends this for the report: Choose a risk threshold, do the analysis, and modify/share your data based upon that analysis.
- Is this a reasonable recommendation given that none of us are doing the statistical analysis routinely?
- David Gutman: Studies that leave out age and sex are not interesting.
- Some TCIA collections are useful even without age and sex.
- HIPAA Safe Harbor is only useful in the US. We are trying to do more than these 18 elements when it comes to de-id.
- In radiology we have traditionally just relied on lists.
- Information can be derived or approximated from images. If pixel data render the data unique, change it to make it less recognizable or delete it. The analysis can lead you to a decision on how to handle this.
- We would be better off recommending this and then over the next several years, hopefully there will be more research into practical ways of doing this.
- Invite selected people from the statistical disclosure community to comment on this and say which tools they use.
- In Europe, GDPR has gotten to the point where you can't share any data. Do we want to go that route?
- Rather than get so extreme, maybe just leave age out of this data, or change it to meet the risk threshold.

- Is there a scalable way to do this?
- Radiology has been immature about this and has not considered existing research into approaches.
- David C. recommends that we read the CAR papers: Canadian Association of Radiologists White Paper on De-Identification of Medical Imaging:
    - Part 1, General Principles
    - Part 2, Practical Considerations
- David G: I realize this is a nuance.. but if we recommend X, and many of the people on this group don't currently do X because it's extremely difficult/nebulous. do we shoot ourselves in the foot? bone density?
- Fred: Hospitals in different countries. Multiple ethics review boards in the same country. Need a threshold that is agreed upon if you are going to de-id anything.
- The risk is finite, so we need to pick a threshold.
- Threshold: Probability of re-id based on threat model. Pick the most conservative one and compute the probability.
- Justin: I'd be very interested to try applying one or two of these automated tools David mentioned against a couple of TCIA datasets to see what happens and help inform the recommendation in the report.
- David C: Yes, let's try this.
- People need to balance utility and risk and insure themselves in the meantime.
- Brian: Utility has zero value to legal people. Risk is always increasing because technology gets better and better.
- Countermeasures which hopefully will keep everything balanced. With released data, unless you are going to pull it back and not release it, the risk always goes up.
- Should we get a guest speaker for the next meeting? Group says yes.

ITEM 8

- For clinical data elements, the general principles in the PHUSE for CDISC SDTM should be applied.
    - Ferran et al https://www.pharmasug.org/proceedings/2015/DS/PharmaSUG-2015-DS10.pdf
    - https://advance.phuse.global/display/WEL/PHUSE+De-identification+Standards
- Ying asked if there is any overlap between these two standards. David says yes but doesn't know if they are conflicting. Someone should do a gap analysis.
- In general, if you're dealing with non-DICOM data, you should apply DICOM elements to it.

ITEM 9

- This is an obvious item that should be stated.
- David is undecided how to deal with stratification. Do people agree that it's fine to stratify images and not check each one?
- Brian checks each one because the metadata is not reliable enough to stratify it. Others agree with looking at everything and not stratifying.

Actions

- David Clunie will recruit a statistician with the right expertise to speak with us at the March 8 meeting.
- David will continue refining the Interim Report.
- All Task Group members are welcome to email David your comments on the report.

March 1, 2022 Meeting

Presentation by Khaled El Emam on Re-identification Risk Measurement - slides, slides with annotations.

- Questions for group were how to pick a threat model, which identifiers to be concerned about, and how to establish a risk threshold for public data release.
- Apply stratification principles to structured data. If you have unstructured data, structure it first.
- Identity disclosure, which is just one type of disclosure but the type most applicable to re-id, is when a person's identity is assigned to a record.
- Trying to measure the risk of verification for a dataset
- Quasi-identifiers are those known by an attacker.
- Delete or encrypt/hash direct identifiers first. What we end up after that is synonymous data.
- For the purposes of re-id risk, we only care about quasi-identifiers.
- A meaningful re-id teaches you something new about the person.
- Attack in two directions - population to sample, sample to population
- Risk is measured by the group size (of 1 = unique)
- Assign a risk value to each record in the dataset.
- To reduce the risk, you can generalize the records and reduce the match rate.
- You can suppress records, remove records, and add noise to reduce the risk of re-id as well.
- generalize - group size gets bigger - risk reduces - maximum (k-anonymity)(public), average (non-public), unicity (proportion of records that are unique in the population)
- You don't want to measure the risk in the data set but measure the risk in the population. The data set is just a sample from the population.
- The group size in the population is the number that's important, but you have to estimate it, since you don't usually have a population registry.
- Once you can estimate the risk properly, you can manage risk in a less conservative way that is still defensive.
- There's no such thing as a probability of zero.
- For releasing public data, a threshold in popular use today is .09. This will give you higher data quality. For particularly sensitive data sets, you would use the more strict threshold of .05.
- risk denominator is not group size in sample but in population
- risk threshold in identifiability spectrum
- privacy-utility trade-off
- data transformations - generalization, suppression, addition of noise, microaggregation
- for non-public data, can add controls (privacy, security, contractual) to deal with residual risk.
- motivated intruder attack-empirical way to evaluate your risk. Commission a white hat attack.
- Two approaches for risk assessment: 1) model-based 2) motivated intruder attack.
- Useful for public data releases. Helps find quasi-identifiers you didn't consider.
- For public data releases, it's harder to release complex data sets and still retain utility.

Two publicly available tools

- SDC Micro (R package) - link - main paper - GUI application paper
- ARX - link - main paper - list of papers

Papers

- motivated intruder attack - Branson et al
- confidence instead of known identity (UK) - Tudor et al

El Emam background and bibliography

- Wikipedia entry with bibliography
- Homepage at uOttawa

Discussion

- Q: Alzheimer's MRI data set. If I heard that an Asian man died of Alzheimer's in Atlanta on a certain date, can I find the person's brain in the data set? What is this called?
- El Emam: You can estimate all of these things. The methods I described can be used to estimate those values.
- Unless your data set includes everyone who ever had Alzheimer's, you don't have the full population.
- Worry that we underestimate populations and destroy data more than we need to.
- El Emam: You have to use your best judgement to come up with these numbers and document the process. In practice, if you go through these methods it's hard to re-id a person. It's not impossible but very difficult.
- We don't know who these people are. There is no registry of who has brain tumors in the US. We have death records.
- El Emam: There's a buffer there. Only in a quarter of cases are you able to validate a match.
- Data quality issues
- Data sets that are published have errors in them
- When you factor in verification of suspected matches plus data quality issues, risk goes down quite a bit.
- Risk is below the threshold in practice.
- How does it work well in practice?
- El Emam: When motivated intruder attacks were done on properly de-identified data, nothing was found.
- El Emam: Weight of evidence is on the pragmatic process. Allows us to release useful data. Model that is too conservative doesn't allow you to release useful data-an exercise in theater. You can make your threshold more strict over time.
- Are there off-the-source tools one can use to do this–estimating population and computing risk?
- How do we determine the real risk? Who would want to find out who an image belongs to in TCIA? How do I define the real risk?
- El Emam: It's a legal requirement and there's motivation by academics and the media. Can build a career.
- All re-id attacks are done by academics and the media.
- There's a risk being a soft target.
- When a dataset it is incrementally increased, such as 100 new individuals added to a dataset of 10,000, do you estimate re-id risk based on the delta or the whole population? It depends how large the delta is. It can be a statistical/methodology argument for estimating population group sizes.

April 12, 2022 Meeting

Interim Report Best Practices And Recommendations Extract as of 20220411

Agenda

- Whole-slide images de-identification goals and issues
- Define project to assess statistical risk of re-identification from images reconstructed as faces
- Continue review of draft best practices document

Discussion

- Introduction to new potential task group member, David Brundage, professor at Cornell.
- Whole-slide images (WSI) are not usually in DICOM format and must be converted.
- Dave Gutman shared slides about protected health information in WSI.
- Most PHI lives in the slide label. Sometimes only a partial label is scanned, so a human might not realize PHI is there, but a machine can detect that PHI.
- The primary image is unlikely to contain PHI.
- Luke Geneslaw also shared some slides about detecting label presence in tissue image scans.
- Trade-off between missing tissue that has cancer on it and bigger files that could have more PHI in the data.
- Leaving data in the slide label causes problems for de-id. There isn't software out there to redaction of pixel data from the tissue sample. Dave Gutman is working on an NCI project to develop it.
- JPEG stores data in 8x8 blocks, so it's possible to remove individual blocks from an image.
- Metadata extraction is unique to format.
- TCIA has a dictionary of private data elements from DICOM.
- Python package: https://github.com/DigitalSlideArchive/tifftools
- Date can be in TIFF times and other data elements defined in the XML, or included as an annotation.
- It's our job to identify which areas need to be mitigated.
- Not all slides are standard formats, like prostate whole mounts.
- David Clunie would like to create a sub-group of people with a special interest and experience in WSI so that they can create content on that subject for the report. Fred Prior, Dave Gutman, David Brundage will join.
- We need a person with significant statistical knowledge who could adapt their knowledge to defacing. Justin suggested someone and will talk to David about it.

- David shared the new version of the task group's de-id report. Please follow the tracked changes offline and if you have any comments on it, please let him know.
- Common stratification based on type–the task group determined that this is not sufficient and we should be looking at everything regardless. Sentence added to report.
- We have not yet defined a best practice on how to score risk. Further research is needed.

June 14, 2022 Meeting

Publications mentioned during the meeting

- TY - RPRT TI - FRVT 2006 and ICE 2006 large-scale results AU - Phillips, P Jonathon AU - Scruggs, W Todd AU - O'Toole, Alice J AU - Flynn, Patrick J AU - Boyer, Kevin W AU - Schott, Cathy L AU - Sharpe, Matthew PY - 2007 PB - National Institute of Standards and Technology CY - Gaithersburg, MD SN - NIST IR 7408 DO - 10.6028/NIST.IR.7408 ER -
  - https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir7408.pdf
  - doi:10.6028/NIST.IR.7408

and

- Schwarz CG, Kremers WK, Lowe VJ, Savvides M, Gunter JL, Senjem ML, Vemuri P, Kantarci K, Knopman DS, Petersen RC, Jack CR Jr; Alzheimer's Disease Neuroimaging Initiative. Face recognition from research brain PET: An unexpected PET problem. Neuroimage. 2022 Jun 3;258:119357. doi: 10.1016/j.neuroimage.2022.119357. Epub ahead of print. PMID: 35660089.
  - https://pubmed.ncbi.nlm.nih.gov/35660089/

Best practices document draft will be ready in the next couple of weeks and will be shared for comment.

Demo of SynthStrip

- Dr. Malte Hoffman's slides (request an accessible version)
- SynthStip prevents facial reconstruction from medical images with pixel-level de-identification.
- Combined effort led by Andrew at the Martinez Center.
- He will present the tool and how it works and then will focus on its evaluation and robustness.
- Skull stripping removes non-brain tissue from CT scan images. The reason we do this is because irrelevant information can appear in downstream analysis algorithms. If we remove these structures, we can increase security. It's a holistic approach to de-id and de-facing.
- Other tools make strong assumptions about the type of image you can use them on. They are dependent on the type of image, like MRI scans. Resolution expectations.
- Implementations of skull stripping from deep learning have limitations too because they train. They would not do as well on new contrast types or new modalities.
- The goal is to get these neural networks to generalize to data that is unseen in training. We suggest to synthesize images of arbitrary characteristics and get these images to train label maps to do the synthesis.
- The SynthStrip team's general process is to start from a set of brain label maps but no images. Given an input label map, they create a new label map by flying a random nonlinear formation to increase spatial variability. Then they sample a right-scale image with a different intensity distribution for every structure in the image, supplying random artifacts such as blurring.
- Since this process is completely randomized, if they were to input the same underlying segmentation map again, they would get a completely different image. They want to encourage their networks to generalize beyond a specific data type.
- For brain extraction, they sample and augment a label map and synthesize an arbitrary image from it. This image is fed into a neural network. Then they compute the ground truth and compare.
- Label maps are only needed for training. If you want to extract the brain from the image, you don't need the label map at test time.
- At evaluation time, we find that SynthStrip performs with robustness and high accuracy on MRI scans, isotropic scans that have fixed license, and diffusing weighted imaging that typically have lower resolution.
- Performance in the presence of pathology? SynthStrip did well against data from TCIA.
- To quantify this, they compiled a database of 600 images and created two metrics. For both metrics, SynthStrip outperforms all of the other baselines they tested. It avoids cross-mislabeling.

- Limitations: 1. SynthStrip does not currently support in-utero brain extraction. 2. SynthStrip is also inherently three-dimensional. 3. It is focused on pre-processing so does not include maximum intensity and so forth.
- Runtime performance is very fast.
- It is a simple command-line utility but there is a standalone version available on the Docker hub.
- They use python and pytorch under the hood.

Discussion of Presentation

- Is SynthStrip data useful in the real world?
- The algorithm needs to prepare to work on skull-stripped data.
- Perspective from radiation oncology: they tested methods and the Karina method was able to maintain what is needed for radiation therapy. Ying wants input from this group on if they test re-id, what would be an acceptable performance metric?

Brian Bialecki on De-identification

- His team at ACR is trying to find a way to release this data publicly.
- He would like to get patient consent to share some identifying data.

- They'd like to see the real data to assess both the real world risk of doing nothing and the real world risk of various mitigation approaches.

Other Discussion

- Skull stripping is not a replacement for for de-facing.
- The value of SynthStrip is not in how well the model performs, but rather how the model is created.
- It's difficult to find things that work across different data sets and modalities. This is something we desperately need.
- SynthStrip does not work on slices, it is fully 3D.
- Access model, registered or restricted, for data with a data use agreement. This is what everyone is converging on.
- Record and track who got the data. But some repositories have no such tracking.

Next Meeting

- Skipping July
- August 9, 2022 at 1 p.m. EST

August 9, 2022 Meeting

Agenda

Applicability of SDC tools to medical image metadata – ARX

Discussion

- Statistical re-identification of radiology and pathology data. Data includes metadata and spreadsheets accompanying the images.
- Review of past discussion:
    - Statistical disclosure control
    - Statistical approaches are mentioned in HIPAA. HIPAA has a privacy rule that is an alternative to the Safe Harbor mechanism
    - Presentation by Dr. El Amam.
    - Estimating re-identification risk and attempting to reduce this.
    - Asked Dr. El Amam which tools exist to help to do this. He said ARX and STC Micro.
- Today David demonstrated the ARX Anonymization Tool, which is a java-based package that runs on any platform. See this tool's YouTube channel at https://www.youtube.com/channel/UCcGAF5nQ_O6ResEF-ivsbVQ/videos
    - David demonstrated how to use this tool, based on his limited understanding of it.
    - Imported a spreadsheet of CPTAC proteomic metadata. This dataset has around 65000 records, so it's large enough to use for this demonstration. This data is already in IDC.
    - This dataset has both actual- and quasi-identifiers.
    - He selected quasi-identifiers.
    - You have to set the sensitivity. You can set whether the data is a quasi-identifier. David set Gender and Age as quasi-identifying.
    - Prosecutor, journalist, and marketer attacker model are shown. Risk shown of how successful they will be. The risk is low for the selected dataset.
    - A prosecutor risk occurs if the adversary can know that the target is in the data set.
    - Distribution of risks in a histogram–prosecutor re-id risk on X axis, records affected on Y axis.
    - ARX is an ptimization tool that, through numerical methods, attempts to optimize changes to the data to reduce the risk and at the same time preserve the utility. It can use more than one privacy model.
    - David tested two models: a K anonymity model with a K of 2 and the average re-identification risk, which is a more complicated model.
    - David provided a generalization pattern by which the tool can aggregate the data as it chooses.
    - He demonstrated creating a hierarchy.
    - After anonymization, there are no more uniques and the quality has not been reduced.
    - The tool makes the data a bit fuzzier without getting rid of data.
    - You can see which rows have been anonymized using the Analyze Utility. Ages are binned or data is omitted according to the model's constraints.
    - Demonstration of adding in race and age. The more quasi-identifying you consider together, the more uniques you get.
    - This tool also has an API.
    - A small data set increases re-identification risk since there are more uniques.
- Adam Taylor will explore this tool for HTAN.
- David is working on drafting the report and hopes to have it done by the next meeting. He will include a section on statistical disclosure control and microdata.

September 13, 2022 Meeting

Agenda

- Review draft of MIDI Task Group report that David Clunie emailed on 9/7/22.

Discussion

- David requests that if anything is missing from/contradictory/contraversial (without sufficient justification) in his draft recommendations document to please tell him. This is meant to be a consensus.
- The team should share with David which sections they would like to focus on if they do not intend to review the whole document.
- The whole team should review the document within a few weeks. David will consolidate comments that the team provides in Word Track Changes.
- David Gutman said that longer is better as far as the document because that means more cases are being considered.
- When the document is mature, Keyvan will share it with the Steering Committee.
- Everyone should start making a list of people they want to share the document with so that they can review it before it is broadly released.

- Suggestions to share the final document (perhaps a short section with link to full document) with academic journals including Journal of Digital Imaging, SPIE, Journal of Cancer Imaging.
- Absence of figures was intentional.
- A summarizing figure or table would be useful to add. Maybe team members have their own figures they could reuse to illustrate certain points.
- DICOM - has anyone encrypted attributes and retained them? May be useful to say there are times when encryption makes sense, but not for a public data set.
- David wrote a long section on facial recognition and asks if the team thinks there is too much emphasis.
- His Dates and Times section is weak and needs more input from knowledgeable people.
- Key Findings section - balance may not have been struck. Regarding statistical methods like demonstrated in last month's meeting, they are not current practice, so David did not want to recommend them as a best practice.
- Keyvan would like to have a workshop in Spring 2023. Would like to run a challenge.
- We can wrap up the Task Group after the publication of the report.
- This Task Group could morph into the program group for the challenge.
- Are there any other topics we'd like to consider? Perhaps a project offline to obtain better statistics about facial reidentification risk.
- Synthetic PHI–jury is out.

Action

All team members should review the draft document updated as of 10/3/22 in Word, tracking their changes and suggestions using Word's Track Changes, and email the document back to David by the week of October 9.

Next Meeting

Tuesday, October 18, 1-2 p.m. EST

October 18, 2022 Meeting

Agenda

- Discuss draft document updated as of 10/3/22.

Discussion

  - Team believes draft report is exceptionally well written.
  - Many comments have been received, including recently many from the TCIA team, which David is working on integrating.
  - David request updated ORC IDs.
  - Carolyn will add the Advisory Group names to the draft.
  - Breach mitigation was suggested as a topic for the document but could be institution-specific and so this potentially large topic will be excluded.
  - In this document, risk means probability instead of probability multiplied by severity as a hazard.
  - Difference between direct and indirect identifier. Direct is like SS number, and example of an indirect one is race. Date of birth or name are not unique and yet most people think of them as direct identifiers. David revised the document to bring clarity to this terminology. You remove direct identifiers and generalize indirect identifiers.
  - Comments on statistical disclosure control. Action after this meeting: look at the first best practice from the perspective of performing an expert statistical analysis or not to see if it would cause trouble for someone or be overstated.
  - Institution as a proxy for location. Can you recover location from the data?
  - Best practice: Do a risk analysis or don't release the data publicly.

Summary of this meeting and next steps:

  - Made changes to two of the best practices during the meeting
  - Hoping to get more feedback by the end of the week.
  - More RT feedback is coming.
  - By end of week, David will accept all tracked changes and start circulating to the advisory group

Action

Provide any additional feedback by the end of this week.

Next Meeting

Tuesday, November 8, 1-2 p.m. EST

November 8, 2022 Meeting

Agenda

- Discuss external reviewers for the *Report of the Medical Image De-Identification (MIDI) Task Group - Best Practices and Recommendations* (draft document updated as of 11/4/22)

Discussion

- External reviewers: adding Marcus Hermann and John Snow Labs.
- A long list of external reviewers is useful because it is a long document and that increases the possibility all of it will be reviewed.

- David will send personal invitations to each reviewer to note their specialty and how they can focus their review.
- We will skip the December meeting and resume in the new year.
- Next step: focus on what we're going to do in the future. Disband, reorganize based on new goals, continue?
- David plans to write an executive summary to submit to journals.

Action

Please feel free to recommend other external reviewers.

Next Meeting

Tuesday, January 11, 2023 at 1 p.m. EST