Medical De-identification & Transcription

Worked examples for AnnoTara AI team training · All names and details are fictional

Colour key: red = personal health information (PHI) to remove `[TOKEN]` = replacement label yellow = looks personal but must STAY

1 · SOAP note de-identification

A SOAP note is a doctor's visit note: **S**ubjective (what the patient says), **O**bjective (measurements and tests), **A**ssessment (diagnosis), **P**lan (treatment). The task: remove all 18 HIPAA identifiers, and keep every medical detail.

Before: original note

Patient: Maria Elena Santos, 54 F | **MRN:** 00458812 | **DOB:** 03/14/1972

Date of visit: 09/21/2026 | **Provider:** Dr. James Whitaker, Mercy Family Clinic, 1420 Oak St, Austin, TX 78701

S: Ms. Santos reports chest tightness for 3 days, worse when climbing stairs at her job at Walmart. Her daughter Anna (phone 512-555-0147) brought her in. History of type 2 diabetes and hypertension.

O: BP 148/92, HR 88, SpO2 97%. EKG: normal sinus rhythm. HbA1c 7.9% (drawn 08/30/2026).

A: Stable angina, rule out coronary artery disease. Type 2 diabetes, not well controlled.

P: Start aspirin 81 mg daily and metoprolol 25 mg twice daily. Stress test at Austin Heart Center on 10/02/2026. Follow up in 2 weeks. Email results to msantos72@gmail.com.

After: de-identified note

Patient: `[NAME]`, 54 F | **MRN:** `[ID]` | **DOB:** `[DATE]`/1972

Date of visit: `[DATE]`/2026 | **Provider:** Dr. `[NAME]`, `[FACILITY]`, `[ADDRESS]`, `[CITY]`, TX `[ZIP]`

S: Ms. `[NAME]` reports chest tightness for 3 days, worse when climbing stairs at her job at `[EMPLOYER]`. Her daughter `[NAME]` (phone `[PHONE]`) brought her in. History of type 2 diabetes and hypertension.

O: BP 148/92, HR 88, SpO2 97%. EKG: normal sinus rhythm. HbA1c 7.9% (drawn `[DATE]`/2026).

A: Stable angina, rule out coronary artery disease. Type 2 diabetes, not well controlled.

P: Start aspirin 81 mg daily and metoprolol 25 mg twice daily. Stress test at `[FACILITY]` on `[DATE]`/2026. Follow up in 2 weeks. Email results to `[EMAIL]`.

Removed

- Names: patient, family, doctor
- All exact dates (the year may stay)
- Record numbers (MRN), phone, email
- Street address, city, ZIP code
- Employer, clinic and hospital names

Kept

- Age 54 (only ages over 89 are hidden)
- State "TX" (only places smaller than a state are removed)
- Vital signs, test results, diagnoses
- Drug names and doses
- Durations such as "3 days", "2 weeks"

Common errors QA must catch

Error	Example	Why it matters
Missed identifier	"Ms. Santos reports..." left in	Privacy breach: the most serious error
Date in another format	"drawn Aug 30" not removed	Dates hide in many formats
Over-redaction	"Start [NAME] 81 mg" (aspirin removed)	Destroys the medical meaning
Wrong label	Phone tagged as [ID]	Breaks the client's data format

2 · Medical audio transcription

Listen to the doctor's dictation and type it out, following the client's style guide.

What the doctor says (audio)

> "This is Doctor Whitaker dictating on Maria Santos, *uh*, date of birth three fourteen seventy-two. Patient is a fifty-four-year-old female presenting with, *um*, three days of substernal chest tightness, exertional... **new paragraph**. Vitals, blood pressure one forty-eight over ninety-two, heart rate eighty-eight... Plan, start aspirin eighty-one milligrams P-O daily, metoprolol tartrate twenty-five milligrams B-I-D. **Period.**"

Finished transcript

> This is Dr. Whitaker dictating on Maria Santos, date of birth 03/14/1972. The patient is a 54-year-old female presenting with 3 days of substernal chest tightness, exertional.
>
> Vitals: Blood pressure 148/92, heart rate 88.
>
> Plan: Start aspirin 81 mg p.o. daily, metoprolol tartrate 25 mg b.i.d.

- Remove fillers ("uh", "um"), and follow spoken commands ("new paragraph", "period").
- Write numbers, units and abbreviations in standard form: 148/92, 81 mg, p.o., b.i.d.
- Spell medical terms correctly: substernal, exertional, metoprolol tartrate.

3 · Transcription QA

Check an AI or first-pass transcript against the audio and fix every error.

> **AI draft (about 80% accurate):** "...presenting with three days of `sub sternal` chest tightness, `exceptional`. ... start `a spring` 81 mg `PO` daily, metoprolol `tart rate` 25 mg `BID`."

AI wrote	Correct	Error type
sub sternal	substernal	Spelling / word split
exceptional	exertional	Wrong word: sounds alike, different meaning
a spring	aspirin	Drug name misheard
tart rate	tartrate	Drug name split
PO / BID	p.o. / b.i.d.	Style guide format

4 · Audio & transcript de-identification

Mark every identifier in the transcript **and** its time in the audio, so those parts can be silenced or beeped.

Start	End	Spoken	Type	Transcript becomes
00:00:02.1	00:00:03.0	"Doctor Whitaker"	Provider name	Dr. [NAME]
00:00:04.4	00:00:05.6	"Maria Santos"	Patient name	[NAME]
00:00:06.0	00:00:08.3	"three fourteen seventy-two"	Date of birth	[DATE]/1972

Result: "This is Dr. `[NAME]` dictating on `[NAME]`, date of birth `[DATE]`/1972...", plus the audio file with those three parts silenced.

The 18 HIPAA Safe Harbor identifiers

1. Names	7. Social Security numbers	13. Device identifiers / serial numbers
2. Places smaller than a state (street, city, county, ZIP*)	8. Medical record numbers	14. Web URLs
3. All dates except the year (birth, admission, discharge, death); ages over 89	9. Health plan beneficiary numbers	15. IP addresses
4. Phone numbers	10. Account numbers	16. Biometric identifiers (finger and voice prints)
5. Fax numbers	11. Certificate / license numbers	17. Full-face photos and similar images
6. Email addresses	12. Vehicle identifiers and plate numbers	18. Any other unique identifying number, characteristic or code

*The first 3 digits of a ZIP code may be kept only when that area has more than 20,000 people. Always follow the client's own guideline when it differs.